PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : C12Q 1/68, G06F 17/30

A2

(11) International Publication Number: **WO 00/18960**

(43) International Publication Date: 6 April 2000 (06.04.00)

(21) International Application Number: PCT/US99/22283

(22) International Filing Date: 24 September 1999 (24.09.99)

(30) Priority Data:
60/101,757 25 September 1998 (25.09.98) US

(71) Applicant: MASSACHUSETTS INSTITUTE OF TECHNOLOGY [US/US]; 77 Massachusetts Avenue, Cambridge, MA 02139 (US).

(72) Inventors: LANDERS, John, E.; 77 Massachusettts Avenue, Cambridge, MA 02139 (US). JORDAN, Barbara; 77 Massachusetts Avenue, Cambridge, MA 02139 (US). HOUSMAN, David, E.; 77 Massachusetts Avenue, Cambridge, MA 02139 (US). CHAREST, Alain; 77 Massachusetts Avenue, Cambridge, MA 02139 (US).

(74) Agent: LOCKHART, Helen, C.; Wolf, Greenfield & Sacks, P.C., 600 Atlantic Avenue, Boston, MA 02210 (US).

(81) Designated States: AU, CA, IL, IS, JP, NO, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: METHODS AND PRODUCTS RELATED TO GENOTYPING AND DNA ANALYSIS

(57) Abstract

The invention encompasses methods and products related to genotyping. The method of genotyping of the invention is based on the use of single nucleotide polymorphisms (SNPs) to perform high throughput genome scans. The high throughput method can be performed by hybridizing SNP allele-specific oligonucleotides and a reduced complexity genome (RCG). The invention also relates to methods of preparing the SNP specific oligonucleotides and RCGs, methods of fingerprinting, determining allele frequency for an SNP, characterizing tumors, generating a genomic classification code for a genome, identifying previously unknown SNPs, and related compositions and kits.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

METHODS AND PRODUCTS RELATED TO GENOTYPING AND DNA ANALYSIS

Field of the Invention

The present invention relates to methods and products associated with genotyping. In particular, the invention relates to methods of detecting single nucleotide polymorphisms and reduced complexity genomes for use in genotyping methods as well as to various methods of genotyping, fingerprinting, and genomic analysis. The invention also relates to products and kits, such as panels of single nucleotide polymorphism allele specific oligonucleotides, reduced complexity genomes, and databases for use in the methods of the invention.

Background of the Invention

Genomic DNA varies significantly from individual to individual, except in identical siblings. Many human diseases arise from genomic variations. The genetic diversity amongst humans and other life forms explains the heritable variations observed in disease susceptibility. Diseases arising from such genetic variations include Huntington's disease, cystic fibrosis, Duchenne muscular dystrophy, and certain forms of breast cancer. Each of these diseases is associated with a single gene mutation. Diseases such as multiple sclerosis, diabetes, Parkinson's, Alzheimer's disease, and hypertension are much more complex. These diseases may be due to polygenic (multiple gene influences) or multifactorial (multiple gene and environmental influences) causes. Many of the variations in the genome do not result in a disease trait. However, as described above, a single mutation can result in a disease trait.

The ability to scan the human genome to identify the location of genes which underlie or are associated with the pathology of such diseases is an enormously powerful tool in medicine and human biology.

Several types of sequence variations, including insertions and deletions, differences in the number of repeated sequences, and single base pair differences result in genomic diversity. Single base pair differences, referred to as single nucleotide polymorphisms (SNPs) are the most frequent type of variation in the human genome (occurring at approximately 1 in 10^3 bases). A SNP is a genomic position at which at least two or more alternative nucleotide alleles occur at a relatively high frequency (greater than 1%) in a population. SNPs are well-suited for studying sequence variation because they are relatively stable (i.e., exhibit low

mutation rates) and because single nucleotide variations can be responsible for inherited traits.

Polymorphisms identified using microsatellite-based analysis, for example, have been used for a variety of purposes. Use of genetic linkage strategies to identify the locations of single Mendelian factors has been successful in many cases (Benomar et al. (1995), *Nat. Genet.*, 10:84-8; Blanton et al. (1991), *Genomics*, 11:857-69). Identification of chromosomal locations of tumor suppressor genes has generally been accomplished by studying loss of heterozygosity in human tumors (Cavenee et al. (1983), *Nature*, 305:779-784; Collins et al. (1996), *Proc. Natl. Acad. Sci. USA*, 93:14771-14775; Koufos et al. (1984), *Nature*, 309:170-172; and Legius et al. (1993), *Nat. Genet.*, 3:122-126). Additionally, use of genetic markers to infer the chromosomal locations of genes contributing to complex traits, such as type I diabetes (Davis et al. (1994), *Nature*, 371:130-136; Todd et al. (1995), *Proc. Natl. Acad. Sci. USA*, 92:8560-8565), has become a focus of research in human genetics.

Although substantial progress has been made in identifying the genetic basis of many human diseases, current methodologies used to develop this information are limited by prohibitive costs and the extensive amount of work required to obtain genotype information from large sample populations. These limitations make identification of complex gene mutations contributing to disorders such as diabetes extremely difficult. Techniques for scanning the human genome to identify the locations of genes involved in disease processes began in the early 1980s with the use of restriction fragment length polymorphism (RFLP) analysis (Botstein et al. (1980), *Am. J. Hum. Genet.*, 32:314-31; Nakamura et al. (1987), *Science*, 235:1616-22). RFLP analysis involves southern blotting and other techniques. Southern blotting is both expensive and time-consuming when performed on large numbers of samples, such as those required to identify a complex genotype associated with a particular phenotype. Some of these problems were avoided with the development of polymerase chain reaction (PCR) based microsatellite marker analysis. Microsatellite markers are simple sequence length polymorphisms (SSLPs) consisting of di-, tri-, and tetra-nucleotide repeats.

Other types of genomic analysis are based on use of markers which hybridize with hypervariable regions of DNA having multiallelic variation and high heterozygosity. The variable regions which are useful for fingerprinting genomic DNA are tandem repeats of a short sequence referred to as a mini satellite. Polymorphism is due to allelic differences in the number of repeats, which can arise as a result of mitotic or meiotic unequal exchanges or by DNA slippage during replication.

The most commonly used method for genotyping involves Weber markers, which are abundant interspersed repetitive DNA sequences, generally of the form $(dC\text{-}dA)_n$ $(dG\text{-}dT)_n$. Weber markers exhibit length polymorphisms and are therefore useful for identifying individuals in paternity and forensic testing, as well as for mapping genes involved in genetic diseases. In the Weber method of genotyping, generally 400 Weber or microsatellite markers are used to scan each genome using PCR. Using these methods, if 5,000 individual genomes are scanned, 2 million PCR reactions are performed (5,000 genomes x 400 markers). The number of PCR reactions may be reduced by multiplexing, in which, for instance, four different sets of primer are reacted simultaneously in a single PCR, thus reducing the total number of PCRs for the example provided to 500,000. The 500,000 PCR mixtures are separated by polyacrylamide gel electrophoresis (PAGE). If the samples are run on a 96-lane gel, 5,200 gels must be run to analyze all 500,000 PCR reaction mixtures. PCR products can be identified by their position on the gels, and the differences in length of the products can be determined by analyzing the gels. One problem with this type of analysis is that "stuttering" tends to occur, causing a smeared result and making the data difficult to interpret and score.

More recent advances in genotyping are based on automated technologies utilizing DNA chips, such as the Affymetrix HuSNP Chip™ analysis system. The HuSNP Chip™ is a disposable array of DNA molecules on a chip (400,000 per half inch square slide). The single stranded DNA molecules bound to the slide are present in an ordered array of molecules having known sequences, some of which are complementary to one allele of a SNP-containing portion of a genome. If the same 5,000 individual genome study described above is performed using the Affymetrix HuSNP Chip™ analysis system, approximately 5,000 gene chips having 1,000 or more SNPs per chip would be required. Prior to the chip scan, the genomic DNA samples would be amplified by PCR in a similar manner to conventional microsatellite genotyping. The gene chip method is also expensive and time-intensive.

Summary Of The Invention

The present invention relates to methods and products for identifying points of genetic diversity in genomes of a broad spectrum of species. In particular, the invention relates to a high throughput method of genotyping of SNPs in a genome (e.g. a human genome) using reduced complexity genomes (RCGs) and, in some exemplary embodiments, using SNP allele specific oligonucleotides (SNP-ASO) and specific hybridization reactions performed, for

example, on a surface. The method of genotyping, in some aspects of the invention, is accomplished by scanning a RCG for the presence or absence of a SNP allele. Using this method, tens of thousands of genomes from one species may be simultaneously assayed for the presence or absence of each allele of a SNP. The methods can be automated, and the results can be recorded using a microarray scanner or other detection/recordation devices.

The invention encompasses several improvements over prior art methods. For instance, a genome-wide scan of thousands of individuals can be carried out at a fraction of the cost and time required by many prior art genotyping methods.

The invention, in one aspect, is a method for detecting the presence of a SNP allele in a genomic sample. The method, in one aspect, includes preparing a RCG from a genomic sample and analyzing the RCG for the presence of the SNP allele. In some aspects, the analysis is performed using a hybridization reaction involving a SNP allele specific oligonucleotide (SNP-ASO) which is complementary to a given allele of the SNP and the RCG. If the allele of the SNP is present in the genomic sample, then the SNP-ASO hybridizes with the RCG.

In some aspects, the method is a method for determining a genotype of a genome, whereby the genotype is identified by the presence or absence of alleles of the SNP in the RCG. In other aspects, the method is a method for characterizing a tumor, wherein the RCG is isolated from a genome obtained from a tumor of a subject and wherein the tumor is characterized by the presence or absence of an allele of the SNP in the RCG.

In other aspects, the method is a method for determining allelic frequency for a SNP, and further comprises determining the number of arbitrarily selected genomes from a population which include each allele of the SNP in order to determine the allelic frequency of the SNP in the population.

In some embodiments, the hybridization reaction is performed on a surface and the RCG or the SNP-ASO is immobilized on the surface. In yet other embodiments, the SNP-ASO is hybridized with a plurality of RCGs in individual reactions.

In other aspects, the method includes performing a hybridization reaction involving a RCG and a surface having a SNP-ASO immobilized thereon, repeating the hybridization with a plurality of RCGs from the plurality of genomes, and determining the genotype based on whether the SNP-ASO hybridizes with at least some of the RCGs.

The RCG may be a PCR-derived RCG or a native RCG. In some embodiments, the

RCG is prepared by performing degenerate oligonucleotide priming-PCR (DOP-PCR) using a degenerate oligonucleotide primer having a tag-$(N)_x$-TARGET nucleotide sequence, wherein the TARGET nucleotide sequence includes at least 7 TARGET nucleotides and wherein x is an integer from 0 to 9, and wherein N is any nucleotide. In various embodiments, the TARGET nucleotide sequence includes 8, 9, 10, 11, or 12 nucleotide residues. In other embodiments, x is an integer from 3 to 9 (e.g. 6, 7, 8, or 9). Preferably, the method of genotyping is performed to determine genotypes more than one locus. In other embodiments, the RCG is prepared by performing DOP-PCR using a degenerate oligonucleotide primer having a tag-$(N)_x$-TARGET nucleotide sequence, wherein the TARGET nucleotide sequence includes fewer than 7 TARGET nucleotide residues and wherein x is an integer from 0 to 9, and wherein N is any nucleotide residue.

The methods can be performed on a support. Preferably, the support is a solid support such as a glass slide, a membrane such as a nitrocellulose membrane, etc.

In yet other embodiments, the RCG is prepared by interspersed repeat sequence-PCR (IRS-PCR), arbitrarily primed-PCR (AP-PCR), adapter-PCR, or multiple primed DOP-PCR. In some aspects of the invention the PCR-generated RCG specifically excludes RCGs prepared by IRS-PCR.

In a preferred embodiment, the methods are useful for determining a genotype associated with or linked to a specific phenotype, and the distinct isolated genomes or RCGs are associated with a common phenotype.

The SNP-ASO used according to the methods of the invention are polynucleotides including one allele of two possible nucleotides at the polymorphic site. In one embodiment, the SNP-ASO is composed of from about 10 to 50 nucleotides. In a preferred embodiment, the SNP-ASO is composed of from about 10 to 25 nucleotides.

According to one embodiment, the SNP-ASO is labeled. The methods can, optionally, also include addition of an excess of non-labeled SNP-ASO in which the polymorphic nucleotide residue corresponds to a different allele of the SNP and which is added during the hybridization step. Additionally, a parallel reaction may be performed wherein the labeling of the two SNP-ASOs is reversed. The label on the SNP-ASO in one embodiment is a radioactive isotope. In this embodiment, the labeled hybridized products on the surface may be exposed to an X-ray film to produce a signal on the film which corresponds to the radioactively labeled hybridization products. In another embodiment, the SNP-ASO is

labeled with a fluorescent molecule. In this embodiment, the labeled hybridized products on the surface may be exposed to an automated fluorescence reader to generate an output signal which corresponds to the fluorescently labeled hybridization products.

According to one embodiment, the RCG is labeled. The label on the RCG in one embodiment is a radioactive isotope. In this embodiment, the labeled hybridized products on the surface may be exposed to an X-ray film to produce a signal on the film which corresponds to the radioactively labeled hybridization products. In another embodiment, the RCG is labeled with a fluorescent molecule. In this embodiment, the labeled hybridized products on the surface may be exposed to an automated fluorescence reader to generate an output signal which corresponds to the fluorescently labeled hybridization products.

In one embodiment, a plurality of different SNP-ASOs are attached to the surface. In another embodiment, the plurality includes at least 500 different SNP-ASOs. In yet another embodiment, the plurality includes at least 1000.

In another embodiment, a plurality of SNP-ASOs are labeled with fluorescent molecules, each SNP-ASO being labeled with a spectrally distinct fluorescent molecule. In various embodiments, the number of spectrally distinct fluorescent molecules is two, three, four, five, six, seven, or eight.

In yet another embodiment, the plurality of RCGs are labeled with fluorescent molecules, each RCG being labeled with a spectrally distinct fluorescent molecule. All of the RCGs having a spectrally distinct fluorescent molecule can be hybridized with a single support. In various embodiments the number of spectrally distinct fluorescent molecules is two, three, four, five, six, seven, or eight.

According to other aspects, the invention encompasses methods for characterizing a tumor by assessing the loss of heterozygosity, determining allelic frequency for a SNP, generating a genomic pattern for an individual genome, and generating a genomic classification code for a genome.

In one aspect, the method for characterizing a tumor includes isolating genomic DNA from tumor samples obtained from a plurality of subjects, preparing a plurality of RCGs from the genomic DNA, performing a hybridization reaction involving a SNP-ASO and the plurality of RCGs (e.g. immobilized on a surface), and identifying the presence of a SNP allele in the genomic DNA based on whether the SNP-ASO hybridizes with at least some of the RCGs in order to characterize the tumor. One or more of the RCGs or one or more of the

SNP-ASOs can be immobilized on a surface.

In another aspect, the invention is a method generating a genomic pattern for an individual genome. The method, in one aspect, includes preparing a plurality of RCGs, analyzing the RCGs for the presence of one or more SNP alleles, and identifying a genomic pattern of SNPs for each RCG by determining the presence or absence therein of SNP alleles. In some embodiments, the analysis involves performing a hybridization reaction involving a panel of SNP-ASOs (e.g. ones which are each complementary to one allele of a SNP), and the plurality of RCGs. The genomic pattern can be identified by determining the presence or absence of a SNP allele for each RCG by detecting whether the SNP-ASOs hybridize with the RCGs. In one embodiment, a plurality of SNP-ASOs are hybridized with the support, and each SNP-ASO of the panel is hybridized with a different support than the other SNP-ASO.

In some embodiments, the genomic pattern is a genomic classification code which is generated from the pattern of SNP alleles for each RCG. In other embodiments, the genomic classification code is also generated from the allelic frequency of the SNPs. In yet other embodiments, the genomic pattern is a visual pattern. The genomic pattern may be in physical or electronic form.

In another aspect, the invention includes is a method for generating a genomic pattern for an individual genome. The method includes identifying a genomic pattern of SNP alleles for each RCG by determining the presence or absence therein of selected SNP alleles.

A method for generating a genomic classification code for a genome is provided in another aspect of the invention. The method includes preparing a RCG, analyzing the RCG for the presence of one or more SNP alleles (e.g. ones of known allelic frequency), identifying a genomic pattern of SNP alleles for the RCG by determining the presence or absence therein of SNP alleles, and generating a genomic classification code for the RCG based on the presence or absence (and, optionally, the allelic frequency) of the SNP alleles. In some embodiments, the analysis involves performing a hybridization reaction involving the RCG and a panel of SNP-ASOs (e.g. corresponding to SNP alleles of known allelic frequency), each of which is complementary to one allele of a SNP. The genomic pattern is identified based on whether each SNP-ASO hybridizes with the RCG.

The method for determining allelic frequency for a SNP, in another aspect, includes preparing a plurality of RCGs from distinct isolated genomes, performing a hybridization reaction involving one RCG and a surface having a SNP-ASO immobilized thereon, repeating

the hybridization with each of the plurality of RCGs, and determining the number of RCGs which include each allele of the SNP in order to determine the allelic frequency of the SNP. In other embodiments the RCGs are immobilized on the surface.

In another aspect, the method for generating a genomic pattern for an individual genome includes preparing a plurality of RCGs, performing a hybridization reaction involving a RCG and a surface having a SNP-ASO immobilized thereon, repeating the hybridization step with each of the plurality of RCGs, and identifying a genomic pattern of SNPs for each RCG by determining the presence therein of SNPs based on whether each SNP-ASO hybridizes with each RCG.

The method for generating a genomic classification code for a genome, in another aspect, includes preparing a RCG, performing a hybridization reaction involving the RCG and a panel of SNP-ASOs (e.g. immobilized on a surface), identifying a genomic pattern of SNPs for the RCG by determining the presence therein of SNPs based on whether each SNP-ASO hybridizes with the RCG, and generating a genomic classification code for the RCG based on the identities of the SNPs which hybridize with the RCG, the identities of the SNPs which do not hybridize with the RCG, and, optionally, also based on the allelic frequency of the SNPs. In one embodiment, each SNP-ASO of the panel is immobilized on a separate surface. In another embodiment, more than one SNP-ASO of the panel is being immobilized on the same surface, each SNP-ASO being immobilized on a distinct area of the surface.

In an embodiment, the genomic classification code is encoded as one or more computer-readable signals on a computer-readable medium

In other aspects of the invention, compositions are provided. According to one aspect, the composition is a plurality of RCGs immobilized on a surface, wherein the RCGs are prepared by a method including the step of performing DOP-PCR using a DOP primer having a tag-$(N)_x$- TARGET nucleotide sequence, wherein the TARGET nucleotide sequence includes at least 7 nucleotide residues, wherein x is an integer from 0 to 9, and wherein N is any nucleotide residue. In various embodiments, the TARGET nucleotide sequence includes 8, 9, 10, 11, or 12 nucleotide residues. In other embodiments, x is an integer from 3 to 9 (e.g. 6, 7, 8 or 9).

According to another aspect, the composition is a panel of SNP-ASOs immobilized on a surface, wherein the SNPs are identified by a method including preparing a set of primers from a RCG, performing PCR using the set of primers on a plurality of isolated genomes to

yield DNA products, isolating and, optionally, sequencing the DNA products, and identifying a SNP based on the sequences of the PCR products. In one embodiment, the plurality of isolated genomes includes at least four isolated genomes.

According to another aspect of the invention, a kit is provided. The kit includes a container housing a set of PCR primers for reducing the complexity of a genome, and a container housing a set of SNP-ASOs. The SNPs which correspond to the SNP-ASOs of the kit are preferably present within a RCG made using the PCR primers of the kit with a frequency of at least 50%.

In one embodiment, the set of PCR primers are primers for DOP-PCR. Preferably, the degenerate oligonucleotide primer has a tag-$(N)_x$-TARGET nucleotide sequence, wherein the TARGET nucleotide sequence includes at least 7 nucleotide residues wherein x is an integer from 0 to 9, and wherein N is any nucleotide residue. In various embodiments, the TARGET nucleotide sequence includes 8, 9, 10, 11, or 12 nucleotide residues. In other embodiments, x is an integer from 3 to 9 (e.g., 6, 7, 8 or 9).

In yet other embodiments, the RCG is prepared by IRS-PCR, AP-PCR, or adapter-PCR.

The SNP-ASOs of the invention are polynucleotides including one of the alternative nucleotides at a polymorphic nucleotide residue of a SNP. In one embodiment, the SNP-ASO is composed of from about 10 to 50 nucleotide residues. In a preferred embodiment the SNP-ASO is composed of from about 10 to 25 nucleotide residues. In another embodiment, the SNP-ASOs are labeled with a fluorescent molecule.

According to yet another aspect of the invention, a composition is provided. The composition includes a plurality of RCGs immobilized on a surface, wherein the RCGs are composed of a plurality of DNA fragments, each DNA fragment including a tag $(N)_x$-TARGET nucleotide, wherein the TARGET nucleotide sequence is identical in all of the DNA fragments of each RCG, wherein the TARGET nucleotidesequence includes at least 7 nucleotide residues, wherein x is an integer from 0 to 9, and wherein N is any nucleotide residue. In various embodiments, the TARGET nucleotide sequence includes 8, 9, 10, 11, or 12 nucleotide residues. In other embodiments, x is an integer from 3 to 9 (e.g. 6, 7, 8, or 9).

In one aspect, the invention is a method for identifying a SNP. The method includes preparing a set of primers from a RCG, wherein the RCG is composed of a first set of PCR products, PCR-amplifying a plurality of isolated genomes using the set of primers to yield a

second set of PCR products, isolating, and optionally, sequencing the PCR products, and identifying a SNP based on the sequences of one or both sets of PCR products. In one embodiment, the plurality of isolated genomes is a pool of genomes. Preferably, the isolated genomes are RCGs. RCGs can be prepared in a variety of ways, but it is preferred, in some aspects, that the RCG is prepared by DOP-PCR.

In one embodiment, the method of preparing the set of primers is performed by at least: preparing a RCG, separating the first set of PCR products into individual PCR products, determining the nucleotide sequence of each end of at least one of the PCR products, and generating primers for use in the subsequent PCR step based on the sequence of the ends of the PCR product(s).

The set of PCR products may be separated by any means known in the art for separating polynucleotides. In a preferred embodiment, the set of PCR products is separated by gel electrophoresis. Preferably, one or more libraries are prepared from segments of the gel containing several PCR products and clones are isolated from the library, each clone including a PCR product from the library. In other embodiments, the set of PCR products is separated by high pressure liquid chromatography or column chromatography.

The RCG used to generate primers or PCR products for identifying SNPs can be prepared by PCR methods. Preferably, the RCG is prepared by performing DOP-PCR using a degenerate oligonucleotide primer having a tag-$(N)_x$-TARGET nucleotide sequence, wherein the TARGET nucleotide sequence includes at least 7 TARGET nucleotide residues wherein x is an integer from 0 to 9, and wherein N is any nucleotide residue. In various embodiments, the TARGET nucleotide sequence includes 8, 9, 10, 11, or 12 nucleotide residues. In other embodiments, x is an integer from 3-9 (e.g. 6, 7, 8, or 9). In other embodiments, the RCG is prepared by performing DOP-PCR using a degenerate oligonucleotide primer having a tag-$(N)_x$-TARGET nucleotide sequence, wherein the TARGET nucleotide sequence includes fewer than 7 TARGET nucleotide residues, wherein x is an integer from 0 to 9, and wherein N is any nucleotide residue.

In yet other embodiments, the RCG is prepared by IRS-PCR, AP-PCR, or adapter-PCR.

In a preferred embodiment of the invention, the set of primers is composed of a plurality of polynucleotides, each polynucleotide including a tag $(N)_x$-TARGET nucleotide

sequence, wherein TARGET is the same sequence in each polynucleotide in the set of primers. The sequence of $(N)_x$ is different in each primer within a set of primers. In some embodiments, the set of primers includes at least 4^3, 4^4, 4^5,4^6, 4^7, 4^8, or 4^9 different primers in the set.

In another aspect, the invention is a method for generating a RCG using DOP-PCR. The method includes the step of performing degenerate DOP-PCR using a degenerate oligonucleotide primer having an $(N)_x$- TARGET nucleotide sequence, wherein the TARGET nucleotide sequence includes at least 7 TARGET nucleotide residues and wherein x is an integer from 0 to 9, and wherein N is any nucleotide residue. In various embodiments the TARGET nucleotide sequence includes 8, 9, 10, 11, or 12 nucleotide residues. In other embodiments, x is an integer from 3 to 9 (e.g. 6, 7, 8, or 9).

According to one embodiment, the tag includes 6 nucleotide residues. Preferably the RCG is used in a genotyping procedure. In other embodiments, the RCG is analyzed to detect a polymorphism. The analysis step may be performed using mass spectroscopy.

In another aspect the invention is a method for assessing whether a subject is at risk for developing a disease. The method includes the steps of using the methods of the invention identify a plurality of SNPs that occur in at least, for example 10% of genomes obtained from individuals afflicted with the disease and determining whether one or more of those SNPs occurs in the subject. In the method the affected individuals are compared with the unaffected individuals. Important information can be generated from the observation that there is a difference between affected and unaffected individuals alone.

In other aspects the invention is a method for identifying a set of one or more SNPs associated with a disease or disease risk. The method includes the steps of preparing individual RCGs obtained from subjects afflicted with a disease, using the same set of primers to prepare each RCG, and comparing the SNP allele frequency identified in those RCGs with the same genetic SNP allele frequency in normal (i.e., non-afflicted) subjects to identify SNP associated with the disease. In other aspects the invention is a method for identifying a set of SNPs randomly distributed throughout the genome. The set of SNPs is used as a panel of genetic markers to perform a genome-wide scan for linkage analysis.

In an embodiment, a computer-readable medium having computer-readable signals stored thereon is provided. The signals define a data structure that one or more data components. Each data component includes a first data element defining a genomic

classification code that identifies a corresponding genome. Each genomic classification code classifies the corresponding genome based one or more single nucleotide polymorphisms of the corresponding genome.

In an optional aspect of this embodiment, the genomic classification code is a unique identifier of the corresponding genome.

In an optional aspect of this embodiment, the genomic classification code is based on a pattern of the single nucleotide polymorphisms of the corresponding genome, where the pattern indicates the presence or absence of each single nucleotide polymorphism.

In another optional aspect of this embodiment, each data component also includes one or more data elements, each data element defining an attributes of the corresponding genome.

Each of the embodiments of the invention can encompass various recitations made herein. It is, therefore, anticipated that each of the recitations of the invention involving any one element or combinations of elements can, optionally, be included in each aspect of the invention.

Brief Description Of The Drawings

Figure 1 is a schematic flow chart depicting a method according to the invention for identifying SNPs.

Figure 2 shows data depicting the process of identifying a SNP: (a) depicts a gel in which inter-Alu PCR genomic DNA products prepared from the 8C primer (which has the nucleotide sequence SEQ ID NO:3) were separated; (b) depicts a gel in which inserts from the library clones were separated; and (c) depicts a filter having two positive or matched clones.

Figure 3 depicts the results of a genotyping and mapping experiment: (a) depicts hybridization results obtained using G allele ASO; (b) depicts hybridization results obtained using A allele ASO; (c) is a pedigree of CEPH family #884 with genotypes indicted from (a) and (b); and (d) is a map of chromosome 3q21-23.

Figure 4 is a schematic flow chart depicting a method according to the invention for detecting SNPs.

Figure 5 is a block diagram of a computer system for storing and manipulating genomic information.

Figure 6A is an example of a record for storing information about a genome and/or genes or SNPs within the genome.

Figure 6B is an example of a record for storing genomic information.

Figure 6C is an example of a record for storing information about genes or SNPs within a genome.

Figure 7 is a flow chart of a method for determining whether genomic information of a sample genome such as SNPs match that of another genome.

Figure 8 depicts results obtained from a hybridization reaction involving RCGs prepared by DOP-PCR and SNP-ASOs immobilized on a surface in a microarray format.

Brief Description Of The Sequences

SEQ. ID. NO. 1 is CAGNNNCTG

SEQ. ID. NO. 2 is TTTTTTTTTTCAG

SEQ. ID. NO. 3 is CTT GCA GTG AGC CGA GATC

SEQ. ID. NO. 4 is CTCGAGNNNNNNAAGCGATG

SEQ ID NO. 5- 697 are nucleotide sequences containing SNPs.

Detailed Description Of The Invention

The invention relates in some aspects to genotyping methods involving detection of one or more single nucleotide polymorphisms (SNPs) in a reduced complexity genome (RCG) prepared from the genome of a subject. The invention includes methods of identifying SNPs associated with a disease or with pre-disposition to a disease. The invention further includes methods of screening RCGs prepared from one or more subjects in a population. Such screening can be used, for example, to determine whether the subject is afflicted with, or is likely to become afflicted with, a disorder, to determine allelic frequencies in the population, or to determine degrees of interrelation among subjects in the population. Additional aspects and details of the compositions, kits, and methods of the invention are described in the following sections.

The invention involves several discoveries which have led to new advances in the field of genotyping. The invention is based on the development of high throughput methods for analyzing genomic diversity. The methods combine use of SNPs, methods for reducing the complexity of genomes, and high throughput screening methods. As discussed in the background of the invention, many prior art methods for genotyping are based on use of hypervariable markers such as Weber markers, which predominantly detect differences in

numbers of repeats. Use of a high throughput SNP analysis method is advantageous in view of the Weber marker system for several reasons. For instance, the results of a Weber analysis system are displayed in the form of a gel, which is difficult to read and must be scored by a professional. The high throughput SNP analysis method of the invention provides a binary result which indicates the presence or absence of the SNP in the sample genome. Additionally, the method of the invention requires significantly less work and is considerably less expensive to perform. As described in the background of the invention, the Weber system requires the performance of 500,000 PCR reactions and use of 5,200 gels to analyze 5,000 genomes. The same study performed using the methods of the invention could be performed without using gels. Additionally, SNPs are not species-specific and therefore the methods of the invention can be performed on diverse species and are not limited to humans. It is more tedious to perform inter-species analysis using Weber markers than using the methods of the invention.

Some prior art methods do use SNPs for genotyping but the high throughput method of the invention has advantages over these methods as well. Affymetrix utilizes a HuSNP Chip™ system having an ordered array of SNPs immobilized on a surface for analyzing nucleic acids. This system is, however, prohibitively expensive for performing large studies such as the 5,000 genome study described above.

The invention is useful for identifying polymorphisms within a genome. Another use for the invention involves identification of polymorphisms associated with a plurality of distinct genomes. The distinct genomes may be isolated from populations which are related by some phenotypic characteristic, familial origin, physical proximity, race, class, etc. In other cases, the genomes are selected at random from populations such that they have no relation to one another other than being selected from the same population. In one preferred embodiment, the method is performed to determine the genotype (e.g. SNP content) of subjects having a specific phenotypic characteristic, such as a genetic disease or other trait. Other uses for the methods of the invention involve identification or characterization of a subject, such as in paternity and maternity testing, immigration and inheritance disputes, breeding tests in animals, zygosity testing in twins, tests for inbreeding in humans and animals, evaluation of transplant suitability, such as with bone marrow transplants, identification of human and animal remains, quality control of cultured cells, and forensic testing such as forensic analysis of semen samples, blood stains, and other biological

materials. The methods of the invention may also be used to characterize the genetic makeup of a tumor by testing for loss of heterozygosity or to determine the allelic frequency of a particular SNP. Additionally, the methods may be used to generate a genomic classification code for a genome by identifying the presence or absence of each of a panel of SNPs in the genome and to determine the allelic frequency of the SNPs. Each of these uses is discussed in more detail herein.

The genotyping methods of the invention are based on use of RCGs that can be reproducibly produced. These RCGs are used to identify SNPs, and can be screened individually for the presence or absence of the SNP alleles.

The invention, in some aspects, is based on the finding that the complexity of the genome can be reduced using various PCR and other genome complexity reduction methods and that RCG's made using such methods can be scanned for the presence of SNPs. One problem with using SNP-ASOs to screen a whole genome (i.e. a genome, the complexity of which has not been reduced) is that the signal to noise (S/N) ratio is high due to the high complexity of the genome and relative frequency of occurrence of a particular SNP-specific sequence within the whole genome. When an entire genome of a complex organism is used as the target for allele-specific oligonucleotide hybridization, the target sequence (e.g. about 17 nucleotide residues) to be detected represents only e.g. approximately 10^8-10^9 1 part in 10^8 of the DNA sample (e.g. for a NP-ASO about 17 nucleotides). It has been discovered, according to the invention, that the complexity of the genome can be reduced in a reproducible manner and that the resulting RCG is useful for identifying the presence of SNPs in the whole genome and for genotyping methods. Reduction in complexity allows genotyping of multiple SNPs following performance of a single PCR reaction, reducing the number of experimental manipulations that must be performed. The RCG is a reliable representation of a specific subfraction of the whole genome, and can be analyzed as though it were a genome of considerably lower complexity.

RCGs are prepared from isolated genomes. An "isolated genome" as used herein is genomic DNA that is isolated from a subject and may include the entire genomic DNA. For instance, an isolated genome may be a RCG, or it may be an entire genomic DNA sample. Genomic DNA is a population of DNA that comprises the entire genetic component of a species excluding, where applicable, mitochondrial and chloroplast DNA. Of course, the methods of the invention can be used to analyze mitochondrial, chloroplast, etc., DNA as

well. Depending on the particular species of the subject, the genomic DNA can vary in complexity. For instance, species which are relatively low on the evolutionary scale, such as bacteria, can have genomic DNA which is significantly less complex than species higher on the evolutionary scale. Bacteria such as *E. coli* have approximately 2.4×10^9 grams per mole of haploid genome, and bacterial genomes having a size of less than about 5 million base pairs (5 megabases) are known. Genomes of intermediate complexity, such as those of plants, for instance, rice, have a genome size of approximately 700-1,000 megabases. Genomes of highest complexity, such as maize or humans, have a genome size of approximately 10^9-10^{11}. Humans have approximately 7.4×10^{12} grams per mole of haploid genome.

A "subject" as used herein refers to any type of DNA-containing organism, and includes, for example, bacteria, viruses, fungi, animals, including vertebrates and invertebrates, and plants.

A "RCG" as used herein is a reproducible fraction of an isolated genome which is composed of a plurality of DNA fragments. The RCG can be composed of random or non-random segments or arbitrary or non-arbitrary segments. The term "reproducible fraction" refers to a portion of the genome which encompasses less than the entire native genome. If a reproducible fraction is produced twice or more using the same experimental conditions the fractions produced in each repetition include at least 50% of the same sequences. In some embodiments the fractions include at least 70%, 80%, 90%, 95%, 97%, or 99% of the same sequences, depending on how the fractions are produced. For instance, if a RCG is produced by PCR another RCG can be generated under identical experimental conditions having at a minimum greater than 90% of the sequences in the first RCG. Other methods for preparing a RCG such as size selection are still considered to be reproducible but often produce less than 99% of the same sequences.

A "plurality" of elements, as used throughout the application refers to 2 or more of the element. A "DNA fragment" is a polynucleotide sequence obtained from a genome at any point along the genome and encompassing any sequence of nucleotides. The DNA fragments of the invention can be generated according to any one of two types mechanisms, and thus there are two types of RCGs, PCR-generated RCGs and native RCGs.

PCR-generated RCGs are randomly primed. That is, each of the polynucleotide fragments in the PCR-generated RCG all have common sequences at or near the 5' and 3' end of the fragment (When a tag is used in the primer, all of the 5' and 3' ends are identical. When

a tag is not used the 5' and 3' ends have a series of N's followed by the TARGET sequence (reading in a 5' to 3' direction). The TARGET sequence is identical in each primer, with the exception of multiple-primed DOP-PCR) but the remaining nucleotides within the fragments do not have any sequence relation to one another. Thus, each polynucleotide fragment in a RCG includes a common 5' and 3' sequence which is determined by the constant region of the primer used to generate the RCG. For instance, if the RCG is generated using DOP-PCR (described in more detail below) each polynucleotide fragment would have near the 5' or 3' end nucleotides that are determined by the "TARGET nucleotide sequence". The TARGET nucleotide sequence is a sequence which is selected arbitrarily but which is constant within a set or subset (e.g. multiple primed DOP-PCR) of primers. Thus, each polynucleotide fragment can have the same nucleotide sequence near the 5' and 3' end arising from the same TARGET nucleotide sequence. In some cases more than one primer can be used to generate the RCG. When more than one primer is used, each member of the RCG would have a 5' and 3' end in common with at least one other member of the RCG and, more preferably, each member of the RCG would have a 5' and 3' end in common with at least 5% of the other members of the RCG. For example, if a RCG is prepared using DOP-PCR with 2 different primers having different TARGET nucleotide sequences, a population containing of four sets of PCR products having common ends could be generated. One set of PCR products could be generated having the TARGET nucleotide sequence of the first primer at or near both the 5' and 3' ends and another set could be generated having the TARGET nucleotide sequence of the second primer at or near both the 5' and 3' ends. Another set of PCR products could be generated having the TARGET nucleotide sequence of the second primer at or near the 5' end and the TARGET nucleotide sequence of the first primer at or near the 3' end. A fourth set of PCR products could be generated having the TARGET nucleotide sequence of the second primer at or near the 3' end and the TARGET nucleotide sequence of the first primer at or near the 5' end. The PCR generated genomes are composed of synthetic DNA fragments.

The DNA fragments of the native RCGs have arbitrary sequences. That is, each of the polynucleotide fragments in the native RCG do not have necessarily any sequence relation to another fragment of the same RCG. These sequences are selected based on other properties, such as size or, secondary characteristics. These sequences are referred to as native RCGs because they are prepared from native nucleic acid preparations rather than being synthesized. Thus they are native-non-synthetic DNA fragments. The fragments of the native RCG may

share some sequence relation to one another (e.g. if produced by restriction enzymes). In some embodiments they do not share any sequence relation to one another.

In some preferred embodiments, the RCG includes a plurality of DNA fragments ranging in size from approximately 200 to 2,000 nucleotide residues. In a preferred embodiment, a RCG includes from 95 to 0.05% of the intact native genome. The fraction of the isolated genome which is present in the RCG of the invention represents at most 90% of the isolated genome, and in preferred embodiments, contains less than 50%, 40%, 30%, 20%, 10%, 5%, or 1% of the genome. A RCG preferably includes between 0.05 and 1% of the intact native genome. In a preferred embodiment, the RCG encompasses 10% or less of an intact native genome of a complex organism.

Genomic DNA can be isolated from a tissue sample, a whole organism, or a sample of cells. Additionally, the isolated genomes of the invention are preferably substantially free of proteins that interfere with PCR or hybridization processes, and are also substantially free of proteins that damage DNA, such as nucleases. Preferably, the isolated genomes are also free of non-protein inhibitors of polymerase function (e.g. heavy metals) and non-protein inhibitors of hybridization when the PCR-generated RCGs are formed. Proteins may be removed from the isolated genomes by many methods known in the art. For instance, proteins may be removed using a protease, such as proteinase K or pronase, by using a strong detergent such as sodium dodecyl sulfate (SDS) or sodium lauryl sarcosinate (SLS) to lyse the cells from which the isolated genomes are obtained, or both. Lysed cells may be extracted with phenol and chloroform to produce an aqueous phase containing nucleic acid, including the isolated genomes, which can be precipitated with ethanol.

Several methods can be used to generate PCR-generated RCG including IRS-PCR, AP-PCR, DOP-PCR, multiple primed PCR, and adaptor-PCR. Hybridization conditions for particular PCR methods are selected in the context of the primer type and primer length to produce to yield a set of DNA fragments which is a percentage of the genome, as defined above. PCR methods have been described in many references, see e.g., US Patent Nos. 5,104,792; 5,106,727; 5,043,272; 5,487,985; 5,597,694; 5,731,171; 5,599,674; and 5,789,168. Basic PCR methods have been described in e.g., Saiki et al., Science, 230: 1350 (1985) and U.S. Pat. Nos. 4,683,195, 4,683,202 (both issued Jul. 18, 1987) and U.S. Pat. No. 4,800,159 (issued Jan. 24, 1989). In some aspects of the invention the PCR-generated RCG specifically excludes RCGs prepared by IRS-PCR.

The PCR methods described herein are performed according to PCR methods well-known in the art. For instance, U.S. Patent No. 5,333,675, issued to Mullis et al. describes an apparatus and method for performing automated PCR. In general, performance of a PCR method results in amplification of a selected region of DNA by providing two DNA primers, each of which is complementary to a portion of one strand within the selected region of DNA. The primer is hybridized to a template strand of nucleic acid in the presence of deoxyribonucleotide triphosphates (dATP, dCTP, dGTP, and dTTP) and a chain extender enzyme, such as DNA polymerase. The primers are hybridized with the separated strands, forming DNA molecules that are single stranded except for the region hybridized with the primer, where they are double stranded. The double stranded regions are extended by the action of the chain extender enzyme (e.g. DNA polymerase) to form an extended double stranded molecule between the original two primers. The double stranded DNA molecules are separated to produce single strands which can then be re-hybridized with the primers. The process is repeated for a number of cycles to generate a series of DNA strands having the same nucleotide sequence between and including the primers.

Chain extender enzymes are well known in the art and include, for example, *E. coli* DNA polymerase I, klenow fragment of *E. coli* DNA polymerase I, T4 DNA polymerase, T7 DNA polymerase, recombinant modified T7 DNA polymerase, reverse transcriptase, and other enzymes. Heat stable enzymes are particularly preferred as they are useful in automated thermal cycle equipment. Heat stable polymerases include, for example, DNA polymerases isolated from *bacillus stearothermophilus* (Bio-Rad), *thermus thermophilous* (finzyme, ATCC number 27634), *thermus* species (ATCC number 31674), *thermus aquaticus* strain TV11518 (ATCC number 25105), *sulfolobus acidocaldarius*, described by Bukhrashuili et al., *Biochem. Biophys. Acta.*, 1008:102-07 (1909), *thermus filiformus* (ATCC number 43280), Taq DNA polymerase, commercially available from Perkin-Elmer-Cetus (Norwalk, Connecticut), Promega (Madison, Wis.) and Stratagene (La Jolla, Calif.), and AmpliTaq™ DNA polymerase, a recombinant thermus equitus Taq DNA polymerase, available from Perkin-Elmer-Cetus and described in U.S. Patent No. 4,889,818.

Preferably, the PCR-based RCG generation methods performed according to the invention are automated and performed using thermal cyclers. Many types of thermal cyclers are well-known in the art. For instance, M.J. Research (Watertown, MA) provides a thermal cycler having a peltier heat pump to provide precise uniform temperature control in the

thermal cyclers; DeltaCycler thermal cyclers from Ericomp (San Diego, CA) also are peltier-based and include automatic ramping control, time/temperature extension programming and a choice of tube or microplate configurations. The RoboCycler™ by Stratagene (La Jolla, CA) incorporates robotics to produce rapid temperature transitions during cycling and well-to-well uniformity between samples; and a particularly preferred cycler, is the Perkin-Elmer Applied Biosystems (Foster City, CA) ABI Prism™ 877 Integrated Thermal cycler, which is operated through a programmable interface that automates liquid handling and thermocycling processes for fluorescent DNA sequencing and PCR reactions. The Perkin-Elmer Applied Biosystems machine is designed specifically for high-throughput genotyping projects and fully automates genotyping steps, including PCR product pooling.

Degenerate oligonucleotide primed-PCR (DOP-PCR) involves use of a single primer set, wherein each primer of the set is typically composed of 3 parts. A DOP-PCR primer as used herein can have the following structure:

5'tag-$(N)_x$-TARGET 3'

The "TARGET" nucleotide sequence includes at least 5 arbitrarily selected nucleotide residues that are the same for each primer of the set. x is an integer from 0 to 9, and N is any nucleotide residue. The value of x is preferably the same for each primer of a DOP-PCR primer sety. In other embodiments, the TARGET nucleotide sequence includes at least 6 or 7 and preferably at least 8, 9, or 10 arbitrarily-selected nucleotides. The tag is optional.

A "TARGET nucleotide" can be used herein is selected arbitrarily. A set of primers is used to generate a particular RCG. Each primer in the set includes the same TARGET nucleotide sequence as the other primers. Of course, sets of primers having different TARGET sequences can be combined.

The "tag", as used herein, is a sequence which is useful for processing the RCG but not necessary. The tag, unlike the other sequences in the primer, does not necessarily hybridize with genomic DNA during the initial round of genomic PCR amplification. In later amplification rounds, the tag hybridizes with PCR, amplified DNA. Thus, the tag does not contribute to the sequence initially recognized by the primer. Since the tag does not participate in the initial hybridization reaction with genomic DNA, but is involved in the primer extension process, the PCR products that are formed (i.e., the reproducible DNA fragments) include the tag sequence. Thus, the end products are DNA fragments that have a sequence identical to a sequence found in the genome except for the tag sequence. The tag is

useful because in later rounds of PCR it allows use of a higher annealing temperature than could otherwise be used with shorter oligonucleotides. The arbitrarily selected sequence is positioned at the 3′ end of the primer. This sequence, although arbitrarily selected, is the same for each primer in a set of DOP-PCR primers. From 0 to 9 nucleotide residues ("N" in the formula above) are located at the 5'-end of the TARGET sequence in the DOP-PCR primers of the invention. Each of these residues can be independently selected from naturally-occurring or artificial nucleotide residues. By way of example, each "N" residue can be an inosine or methylcytosine residue. In the formula, "x" is an integer that can be from 0 to 9, and is preferably from 3 to 9 (e.g. 3, 4, 5, 6, 7, 8, or 9). Each set of DOP-PCR primers of the invention can thus contain up to 4^x unique primers (i.e., 1, 4, 16, 64..., 262144 primers for x = 0, 1, 2, 3,..., 9). Finally, a base pair tag can be positioned at the 5′ end of the primer. This tag can optionally include a restriction enzyme site. In general, inclusion of a tag sequence in the DOP-PCR primers of the invention is preferred, but not necessary.

The initial rounds of DOP-PCR are preferably performed at a low temperature given that the specificity of the reaction will be determined by only the 3' TARGET nucleotide sequence. A slow ramp time during these cycles ensures that the primers do not detach from the template before being extended. Subsequent rounds are carried out at a higher annealing temperature because in the subsequent rounds the 5′ end of the DOP-PCR primer (the tag) is able to contribute to the primer annealing. A PCR cycle performed under low stringency hybridization conditions generally is from about 35°C to about 55°C.

Because DOP-PCR involves a randomly chosen sequence, the resultant PCR products are generated from genome sequences arbitrarily distributed throughout the genome and will generally not be clustered within specific sites of the genome. Additionally, creation of new sets of DOP-PCR-amplified DNA fragments can be easily accomplished by changing the sequence, length, or both, of the primer. RCGs having greater or lesser complexity can be generated by selecting DOP-PCR primers having shorter or longer, respectively, TARGET and $(N)_x$ nucleotide sequences. This approach can also be used with multiple DOP-PCR primers such as in the "multiple-primed DOP-PCR" method (described below). Finally, use of arbitrarily chosen sequences of DOP-PCR is useful in many species because the arbitrarily-selected sequences are not species-specific, as with some forms of PCR which require use of a specific known sequence.

Another method for generating a PCR-generated RCG involves interspersed repeat

sequence PCR (IRS-PCR). Mammalian chromosomes include both repeated and unique sequences. Some of the repeated sequences are short interspersed repeated sequences (IRS's) and others are long IRS's. One major family of short IRS's found in humans includes Alu repeat sequences. Amplification using a single Alu primer will occurs whenever two Alu elements lie in inverted orientation to each other on opposite strands. There are believed to be approximately 900,000 Alu repeats in a human haploid genome. Another type of IRS sequence is the L1 element (most common is L1Hs) which is present in 10^4-10^5 copies in a human genome. Because the L1 sequence is expressed less abundantly in the genome than the Alu sequence, fewer amplification products are produced upon amplification using an L1 primer. In IRS-PCR, a primer which has homology to a repetitive sequence present on opposite strands within the genome of the species to be analyzed is used. When two repeat elements having the primer sequence are present in a head-to-head fashion within a limited distance (approximately 2000 nucleotide residues), the inter-repeat sequence can be amplified. The method has the advantage that the complexity of the resulting PCR products can be controlled by how homologous the primer chosen is with the repeat consensus (that is, the more homologous the primer is with the repeat consensus sequence, the more complex the PCR product will be).

In general, an IRS-PCR primer has a sequence wherein at least a portion of the primer is homologous with (e.g. 50%, 75%, 90%, 95% or more identical to) the consensus nucleotide sequence of an IRS of the subject.

In mammalian genomes, small interspersed repeat sequences (SINES) are present in extremely high copy number and are often configured such that a single copy sequence of between 500 nucleotide residues and 1000 nucleotide residues is situated between two repeats which are oriented in a head-to-head or tail-to-tail manner. Genomic DNA sequences having this configuration are substrates for Alu PCR in human DNA and B1 and B2 PCR in the mouse. The precise number of products which are represented in a specific Alu, B1, or B2 PCR reaction depends on the choice of primer used for the reaction. This variation in product complexity is due to the variation in sequence among the large number of representative sequences of the IRS family in each species. A detailed study of this variation was described by Britten (Britten, R.J. (1994), *Proc. Natl. Acad. Sci. USA*, 91:5992-5996). In the Britten study, the sequence variation for each nucleotide residue of the Alu consensus sequence was analyzed for 1574 human Alu sequences. The complexity of Alu PCR products generated by

amplification using a given Alu PCR primer can be predicted to a significant extent based on the degree to which the nucleotide sequence of the primer matches consensus nucleotide sequences. As a general rule, Alu PCR products become progressively less complex as the primer sequence diverges from the Alu consensus. Because two hybridized primers are required at each site for which Alu PCR is to be accomplished, it is predictable that linear variation and the number of genomic sites to which a primer may bind will be reflected in the complexity of PCR products, which is roughly proportional to the square of primer binding efficiency. This prediction conforms to experimental results, permitting synthesis of Alu PCR products having a wide range of product complexity values. Therefore, when it is desirable to reduce the number of PCR products obtained using Alu PCR, the primer sequence should be designed to diverge by a predictable amount from the Alu consensus sequence.

Another method for generating a RCG involves arbitrarily primed PCR (AP-PCR). AP-PCR utilizes short oligonucleotides as PCR primers to amplify a discrete subset of portions of a high complexity genome. For AP-PCR, the primer sequence is arbitrary and is selected without knowledge of the sequence of the target nucleic acids to be amplified. The arbitrary primer is generally 50-60% G+C. The AP-PCR method is similar to the DOP-PCR method described above, except that the AP-PCR primer consists of only the arbitrarily-selected nucleotides and not the 5′ flanking degenerate residues or the tag (i.e. N_x residue described for the DOP-PCR primers). The genome may be primed using a single arbitrary primer or a combination of two or more arbitrary primers, each having a different, but optionally related, sequence.

AP-PCR is performed under low stringency hybridization conditions, allowing hybridization of the primer with targets with which the primer can exhibit a substantial degree of mismatching. A PCR cycle performed under low stringency hybridization conditions generally is from about 35°C to about 55°C. Mismatches refer to non complementary nucleotide bases in the primer, relative to the template with which it is hybridized.

AP-PCR methods have been used previously in combination with gel electrophoresis to determine genotypes. AP-PCR products are generationally fractionated on a high resolution polyacrylamide gel, and the presence or absence of specific bands is used to genotype a specific locus. In general, the difference between the presence and absence of a band is a consequence of a single nucleotide DNA sequence difference in one of the primer binding sites for a given single copy sequence.

The product complexity obtained using a given primer or primer set can be determined by several methods. For instance, the product complexity can be determined using PCR amplification of a panel of human yeast artificial chromosome (YAC) DNA samples from a CEPH 1 library. These YACs each carry a human DNA segment approximately 300-400 kilobase pairs in length. Product complexity for each primer set can be inferred by comparing the number of bands produced per YAC when analyzed on agarose gel with an IRS-PCR product of known complexity. Additionally, for products of relatively low complexity, electrophoresis on polyacrylamide gels can establish the product complexity, compared to a standard. Alternatively, an effective way to estimate the complexity of the product is to carry out a reannealing reaction using resistance to S1 nuclease-catalyzed degradation to determine the rate of reannealing of internally labeled, denatured, double-stranded DNA product. Comparison with reannealing rates of standards of known complexity permits accurate estimation of product complexity. Each of these three methods may be used for IRS PCR. The second and third methods are best for AP-PCR and DOP-PCR which, unlike IRS-PCR, will not selectively amplify human DNA from a crude YAC DNA preparation.

The complexity of PCR products generated by AP-PCR can be regulated by selecting the primer sequence length, the number of primers in a primer set, or some combination of these. By choosing the appropriate combination, AP-PCR may also be used to reduce the complexity of a genome for SNP identification and genotyping, as described herein. AP-PCR markers are different from Alu PCR primers, have a different genomic distribution, and can therefore complement an IRS-PCR genome complexity-reducing method. The methods can be used in combination to produce complementary information from genome scans.

One PCR method for preparing RCGs is an adapter-linker amplification PCR method (previously described in e.g., Saunders et al., Nuc. Acids Res., 17 9027 (1990); Johnson, Genomics, 6: 243 (1990) and PCT Application WO90/00434, published Aug. 9, 1990. In this method, genomic DNA is digested using a restriction enzyme, and a set of linkers is ligated onto the ends of the resulting DNA fragments. PCR amplification of genomic DNA is accomplished using a primer which can bind with the adapter linker sequence. Two possible variations of this procedure which can be used to limit genome complexity are (a) to use a restriction enzyme which produces a set of fragments which vary in length such that only a subset (e.g. those smaller than a PCR-amplifiable length) are amplified; and (b) to digest the genomic DNA using a restriction enzyme that produces an overhang of random nucleotide

sequence (e.g., AlwN1 recognizes CAGNNNCTG; SEQ ID NO: 1) and cleaves between NNN and CTG). Adapters are constructed to anneal with only a subset of the products. For example, in the case of AlwN1, adapters having a specific 3 nucleotide residue overhang (corresponding to the random 3 base pair sequence produced by the restriction enzyme digestion) would be used to yield (4^3) 64-fold reduction in complexity. Fragments which have an overhang sequence complementary to the adapter overhang are the only ones which are amplified.

Another method for generating RCGs is based on the development of native RCGs. Several methods can be used to generate native RCGs, including DNA fragment size selection, isolating a fraction of DNA from a sample which has been denatured and reannealed, pH-separation, separation based on secondary structure, etc.

Size selection can be used to generate a RCG by separating polynucleotides in a genome into different fractions wherein each fraction contains polynucleotides of an approximately equal size. One or more fractions can be selected and used as the RCG. The number of fractions selected will depend on the method used to fragment the genome and to fractionate the pieces of the genome, as well as the total number of fractions. In order to increase the complexity of the RCG, more fractions are selected. One method of generating a RCG involves fragmenting a genome into arbitrarily size pieces and separating the pieces on a gel (or by HPLC or another size fractionation method). A portion of the gel is excised, and DNA fragments contained in the portion are isolated. Typically, restriction enzymes can be used to produce DNA fragments in a reproducible manner.

Separation based on secondary structure can be accomplished in a manner similar to size selection. Different fractions of a genome having secondary structure can be separated on a gel. One or more fractions are excised from the gel, and DNA fragments are isolated therefrom.

Another method for creating a native RCG involves isolating a fraction of DNA from a sample which has been denatured and reannealed. A genomic DNA sample is denatured, and denatured nucleic acid molecules are allowed to reanneal under selected conditions. Some conditions allow more of the DNA to be reannealed than other conditions. These conditions are well known to those of ordinary skill in the art. Either the reannealed or the remaining denatured fractions can be isolated. It is desirable to select the smaller of these two fractions in order to generate RCG. The reannealing conditions used in the particular reaction

determine which fraction is the smaller fraction. Variations of this method can also be used to generate RCGs. For instance, once a portion of the fraction is allowed to reanneal, the double stranded DNA may be removed (e.g., using column chromatography), the remaining DNA can then be allowed to partially reanneal, and the reannealed fraction can be isolated and used. This variation is particularly useful for removing repetitive elements of the DNA, which rapidly reanneal.

The amount of isolated genome used in the method of preparing RCGs will vary, depending on the complexity of the initial isolated genome. Genomes of low complexity, such as bacterial genomes having a size of less than about 5 million base pairs (5 megabases), usually are used in an amount from approximately 10 picograms to about 250 nanograms. A more preferred range is from 30 picograms to about 7.5 nanograms, and even more preferably, about 1 nanogram. Genomes of intermediate complexity, such as plants (for instance, rice, having a genome size of approximately 700-1,000 megabases) can be used in a range of from approximately 0.5 nanograms to 250 nanograms. More preferably, the amount is between 1 nanogram and 50 nanograms. Genomes of highest complexity (such as maize or humans, having a genome size of approximately 3,000 megabases) can be used in an amount from approximately 1 nanogram to 250 nanograms (e.g. for PCR).

In addition to the DOP-PCR methods described above, PCR-generated RCGs can be prepared using DOP-PCR involving multiple primers, which is referred to herein as "multiple-primed-DOP-PCR". Multiple-primed-DOP-PCR involves the use of at least two primers which are arranged similarly to the single primers discussed above and are typically composed of 3 parts. A multiple-primed-DOP-PCR primer as used herein has the following structure:

$$\text{tag-}(N)_x\text{-TARGET}_2$$

The $TARGET_2$ nucleotide sequence includes at least 5, and preferably at least 6, TARGET nucleotide residues, x is an integer from 0-9, and N is any nucleotide residue.

The sequence chosen arbitrarily and positioned at the 3′ end of the primer can be manipulated in multiple-primed-DOP-PCR to produce a different end product than for DOP-PCR because use of two or more sets of primers adds another level of diversity, thus producing a RCG or amplified genome, depending on the primers chosen. Each of the at least two sets of primers of multiple-primed-DOP-PCR has a different TARGET sequence. Similar to the single primer of DOP-PCR a set of primers is generated for each of the at least two primers and, every primer within a single set has the same TARGET sequence as the other

primers of the set. This TARGET sequence is flanked at its 5' end by 0 to 9 nucleotide residues ("N"s). The set of N's will differ from primer to primer within a set of primers. A set of primers may include up to 4^x different primers, each primer having a unique $(N)_x$ sequence. Finally a tag can be positioned at the 5' end.

In other aspects of the invention, methods for identifying SNPs can be performed using RNA genomes rather than RCGs. RNA genomes differ from RCGs in that they are generated from RNA rather than from DNA. An RNA genome can be, for instance, a cDNA preparation made by reverse transcription of RNA obtained from cells of a subject (e.g. human ovarian carcinoma cells). Thus, an RNA genome can be composed of DNA sequences, as long as the DNA is derived from RNA. RNA can also be used directly.

The genotyping and other methods of the invention can also be performed using a RNA genotyping method. This method involves use of RNA, rather than DNA, as the source of nucleic acid for genotyping. In this embodiment, RNA is reverse transcribed (e.g. using a reverse transcriptase) to produce cDNA for use as an RNA genome. The RNA method has at least one advantage over DNA-based methods. SNPs in coding regions (cSNPs) are more likely to be directly involved in detectable phenotypes and are thus more likely to be informative with regard to how such phenotypes can be affected. Furthermore, since this method can require only a reverse transcription step, it is amenable to high-throughput analysis. In a preferred embodiment, a reverse transcriptase primer which only binds a subset of RNA species (e.g. a dT primer having a 3-base anchor, e.g. TTTTTTTTTT CAG; SEQ ID NO: 2) is used to further reduce RNA genome complexity (48-fold using the dt-3base anchor primer). In the RNA-genotyping method of the invention the RNA/cDNA sample can be attached to a surface and hybridized with a SNP-ASO.

In another aspect, the invention includes a method for identifying a SNP. Genomic fragments which include SNPs can be prepared according to the invention by preparing a set of primers from a RCG (e.g., a RCG is composed of a set of PCR products), performing PCR using the set of primers to amplify a plurality of isolated genomes to produce DNA products, and identifying SNPs included in the DNA products. The presence of a SNP in the DNA product can be identified using methods such as direct sequencing, i.e. using dideoxy chain termination or Maxam Gilbert (see e.g., Sambrook et al, "Molecular Cloning: A Laboratory Manual," Cold Spring Harbor Laboratory, 1989, New York; or Zyskind et al., Recombinant DNA Laboratory Manual, Acad. Press, 1988), denaturing gradient gel electrophoresis to

identify different sequence dependent melting properties and electrophoretic migration of SNPs containing DNA fragments (see e.g., Erlich, ed., PCR Technology, Principles and applications for DNA Amplification, Freeman and Co., NY, 1992), and conformation analysis to differentiate sequences based on differences in electrophoretic migration patterns of single stranded DNA products (see e.g., Orita et al., Proc. Nat. Acad.. Sci. 86, 2766-2770, 1989). In preferred embodiments, the SNPs are identified based on the sequences of the polymerase chain reaction products identified using sequencing methods.

A "single nucleotide polymorphism" or "SNP" as used herein is a single base pair (i.e., a pair of complementary nucleotide residues on opposite genomic strands) within a DNA region wherein the identities of the paired nucleotide residues vary from individual to individual. At the variable base pair in the SNP, two or more alternative base pairings occur at a relatively high frequency (greater than 1%) in a subject, (e.g. human) population.

A "polymorphic region" is a region or segment of DNA the nucleotide sequence of which varies from individual to individual. The two DNA strands which are complementary to one another except at the variable position are referred to as alleles. A polymorphism is allelic because some members of a species have one allele and other members have a variant allele and some have both. When only one variant sequence exists, a polymorphism is referred to as a diallelic polymorphism. There are three possible genotypes in a diallelic polymorphic DNA in a diploid organism. These three genotypes arise because it is possible that a diploid individual's DNA may be homozygous for one allele, homozygous for the other allele, or heterozygous (i.e. having one copy of each allele). When other mutations are present, it is possible to have triallelic or higher order polymorphisms. These multiple mutation polymorphisms produce more complicated genotypes.

SNPs are well-suited for studying sequence variation because they are relatively stable (i.e. they exhibit low mutation rates) and because it appears that SNPs can be responsible for inherited traits. These properties make SNPs particularly useful as genetic markers for identifying disease-associated genes. SNPs are also useful for such purposes as linkage studies in families, determining linkage disequilibrium in isolated populations, performing association analysis of patients and controls, and loss of heterozygosity studies in tumors.

An exemplary method for identifying SNPs is presented in the Examples below. Briefly, DOP-PCR is performed using genomic DNA obtained from an individual. The products are separated on an agarose gel. The products are separated by approximate length

into approximately 8 segments having sizes of about 400-1000 base pairs, and libraries are made from each of the segments. This approach prevents domination of the library by one or two abundant products. Plasmid DNA is isolated from individual colonies containing portions of the library. Inserts are isolated and the ends of the inserts are sequenced using vector primers. A new set of primers is then synthesized based on these insert sequences to allow PCR to be performed using RCG obtained from one or more individuals or from a pool of individuals. The DNA products generated by the PCR are sequenced and inspected for the presence of two nucleotide residues at one location, an indication that a polymorphism exists at that position within one of the alleles.

A "primer" as used herein is a polynucleotide which hybridizes with a target nucleic acid with which it is complementary and which is capable of acting as an initiator of nucleic acid synthesis under conditions for primer extension. Primer extension conditions include hybridization between the primer and template, the presence of free nucleotides, a chain extender enzyme, e.g., DNA polymerase, and appropriate temperature and pH.

In preferred embodiments, a set of primers is prepared by at least the following steps: preparing a RCG, composed of a set of PCR products, separating the set of PCR products into individual PCR products, determining the sequence of each end of at least one of the PCR products, and generating the set of primers for use in the subsequent PCR step based on the sequence of the ends of the insert(s).

A "set of PCR products", as used herein, is a plurality of synthetic polynucleotide sequences, each polynucleotide sequence being different from one another except for a stretch of nucleotides in the 5' and 3' regions of the polynucleotides which are identical in each polynucleotide. These regions correspond to the primers used to generate the RCG and the sequence in these regions varies depending on what primer is used. When a DOP PCR primer is used, the sequence that varies in each primer preferably has a sequence N_x, wherein x is 5-12 and N is any nucleotide. A set of DNA products is different from a "set of PCR products" as used herein and refers to DNA generated by PCR using specific primers which amplify a specific locus.

Once the sequence of a primer is known, the primer may be purified from a nucleic acid preparation which includes, it or it may be prepared synthetically. For instance, nucleic acid fragments may be isolated from nucleic acid sequences in genomes, plasmids, or other vectors by site-specific cleavage, etc. Alternatively, the primers may be prepared by *de novo*

chemical synthesis, such as by using phosphotriester or phosphodiester synethetic methods, such as those described in U.S. Patent No. 4,356,270; Itakura et al. (1989), *Ann. Rev. Biochem.*, 53:323-56; and Brown et al. (1979), *Meth. Enzymol.*, 68:109. Primers may also be prepared using recombinant technology, such as that described in Sambrook, "Molecular Cloning: A Laboratory Manual," Cold Spring Harbor Laboratory, p. 390-401 (1982).

The term "nucleotide residue" refers to a single monomeric unit of a nucleic acid such as DNA or RNA. The term "base pair" refers to two nucleotide residues which are complementary to one another and are capable of hydrogen bonding with one another. Traditional base pairs are between G:C and T:A. The letters G, C, T, U and A refer to (deoxy)guanosine, (deoxy)cytidine, (deoxy)thymidine, uridine, and (deoxy)adenosine, respectively. The term "nucleic acids" as used herein refers to a class of molecules including single stranded and double stranded deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and polynucleotides. Nucleic acids within the scope of the invention include naturally occurring and synthetic nucleic acids, nucleic acid analogs, modified nucleic acids, nucleic acids containing modified nucleotides, modified nucleic acid analogs, and mixtures of any of these.

SNPs identified or detected in the genotyping methods described herein can also be identified by other methods known in the art. Many methods have been described for identifying SNPs. (see e.g. W095/12607, Bostein, et al., *Am. J. Hum. Genet*,.32:314-331 (1980), etc.). In some embodiments, it is preferred that SNPs be identified using the same method that will subsequently be used for genotype analysis.

As discussed briefly above, the SNPs and RCGs of the invention are useful for a variety of purposes. For instance, SNPs and RCGs are useful for performing genotyping analysis; for identification of a subject, such as in paternity or maternity testing, in immigration and inheritance disputes, in breeding tests in animals, in zygosity testing in twins, in tests for inbreeding in humans and animals; in evaluation of transplant suitability such as with bone marrow transplants; in identification of human and animal remains; in quality control of cultured cells; in forensic testing such as forensic analysis of semen samples, blood stains, and other biological materials; in characterization of the genetic makeup of a tumor by testing for loss of heterozygosity; in determining the allelic frequency of a particular SNP; and in generating a genomic classification code for a genome by identifying the presence or absence of each of a panel of SNPs in the genome of a subject and

optionally determining the allelic frequency of the SNPs.

A preferred use of the invention is in a high throughput method of genotyping. "Genotyping" is the process of identifying the presence or absence of specific genomic sequences within genomic DNA. Distinct genomes may be isolated from individuals of populations which are related by some phenotypic characteristic, by familial origin, by physical proximity, by race, by class, etc. in order to identify polymorphisms (e.g. ones associated with a plurality of distinct genomes) which are correlated with the phenotype family, location, race, class, etc. Alternatively, distinct genomes may be isolated at random from populations such that they have no relation to one another other than their origin in the population. Identification of polymorphisms in such genomes indicates the presence or absence of the polymorphisms in the population as a whole, but not necessarily correlated with a particular phenotype.

Although genotyping is often used to identify a polymorphism associated with a particular phenotypic trait, this correlation is not necessary. Genotyping only requires that a polymorphism, which may or may not reside in a coding region, is present. When genotyping is used to identify a phenotypic characteristic, it is presumed that the polymorphism affects the phenotypic trait being characterized. A phenotype may be desirable, detrimental, or, in some cases, neutral.

Polymorphisms identified according to the methods of the invention can contribute to a phenotype. Some polymorphisms occur within a protein coding sequence and thus can affect the protein structure, thereby causing or contributing to an observed phenotype. Other polymorphisms occur outside of the protein coding sequence but affect the expression of the gene. Still other polymorphisms merely occur near genes of interest and are useful as markers of that gene. A single polymorphism can cause or contribute to more than one phenotypic characteristic and, likewise, a single phenotypic characteristic may be due to more than one polymorphism. In general multiple polymorphisms occurring within a gene correlate with the same phenotype. Additionally, whether an individual is heterozygous or homozygous for a particular polymorphism can affect the presence or absence of a particular phenotypic trait.

Phenotypic correlation is performed by identifying an experimental population of subjects exhibiting a phenotypic characteristic and a control population which do not exhibit that phenotypic characteristic. Polymorphisms which occur within the experimental population of subjects sharing a phenotypic characteristic and which do not occur in the

control population are said to be polymorphisms which are correlated with a phenotypic trait. Once a polymorphism has been identified as being correlated with a phenotypic trait, genomes of subjects which have potential to develop a phenotypic trait or characteristic can be screened to determine occurrence or non-occurrence of the polymorphism in the subjects' genomes in order to establish whether those subjects are likely to eventually develop the phenotypic characteristic. These types of analyses are generally carried out on subjects at risk of developing a particular disorder such as Huntington's disease or breast cancer.

A phenotypic trait encompasses any type of genetic disease, condition, or characteristic, the presence or absence of which can be positively determined in a subject. Phenotypic traits that are genetic diseases or conditions include multifactorial diseases of which a component may be genetic (e.g. owing to occurrence in the subject of a SNP), and predisposition to such diseases. These diseases include such as, but not limited to, asthma, cancer, autoimmune diseases, inflammation, blindness, ulcers, heart or cardiovascular diseases, nervous system disorders, and susceptibility to infection by pathogenic microorganisms or viruses. Autoimmune diseases include, but are not limited to, rheumatoid arthritis, multiple sclerosis, diabetes, systemic lupus, erythematosus and Grave's disease. Cancers include, but are not limited to, cancers of the bladder, brain, breast, colon, esophagus, kidney, hematopoietic system eg. leukemia, liver, lung, oral cavity, ovary, pancreas, prostate, skin, stomach, and uterus. A phenotypic characteristic includes any attribute of a subject other than a disease or disorder, the presence or absence of which can be detected. Such characteristics can, in some instances, be associated with occurrence of a SNP in a subject which exhibits the characteristic. Examples of characteristics include, but are not limited to, susceptibility to drug or other therapeutic treatments, appearance, height, color (e.g. of flowering plants), strength, speed (e.g. of race horses), hair color, etc. Many examples of phenotypic traits associated with genetic variation have been described, see e.g., US Patent No. 5,908,978 (which identifies association of disease resistance in certain species of plants associated with genetic variations) and US Patent No. 5,942,392 (which describes genetic markers associated with development of Alzheimer's disease).

Identification of associations between genetic variations (e.g. occurrence of SNPs) and phenotypic traits is useful for many purposes. For example, identification of a correlation between the presence of a SNP allele in a subject and the ultimate development by the subject of a disease is particularly useful for administering early treatments, or instituting lifestyle

changes (e.g., reducing cholesterol or fatty foods in order to avoid cardiovascular disease in subjects having a greater-than-normal predisposition to such disease), or closely monitoring a patient for development of cancer or other disease. It may also be useful in prenatal screening to identify whether a fetus is afflicted with or is predisposed to develop a serious disease. Additionally, this type of information is useful for screening animals or plants bred for the purpose of enhancing or exhibiting of desired characteristics.

One method for determining a genotype associated with a plurality of genomes is screening for the presence or absence of a SNP in a plurality of RCGs. For example, such screening may be performed using a hybridization reaction including a SNP-ASO and the RCGs. Either the SNP-ASO or the RCGs can, optionally be immobilized on a surface. The genotype is determined based on whether the SNP-ASO hybridizes with at least some of the RCGs. Other methods for determining a genotype involve methods which are not based on hybridization, including, but not limited to, mass spectrometric methods. Methods for performing mass spectrometry using nucleic acid samples have been described. See e.g., US Patent No. 5,885,775. The components of the RCG can be analyzed by mass spectrometry to identify the presence or absence of a SNP allele in the RCG.

A "SNP-ASO", as used herein, is an oligonucleotide which includes one of two alternative nucleotides at a polymorphic site within its nucleotide sequence. In some embodiments, it is preferred that the oligonucleotide include only a single mismatched nucleotide residue namely the polymorphic residue, relative to an allele of a SNP. In other cases, however, the oligonucleotide may contain additional nucleotide mismatches such as neutral bases or may include nucleotide analogs. This is described in more detail below. In preferred embodiments, the SNP-ASO is composed from about 10 to 50 nucleotide residues. In more preferred embodiments, it is composed of from about 10 to 25 nucleotide residues.

Oligonucleotides may be purchased from commercial sources such as Genosys, Inc., Houston, Texas or, alternatively, may be synthesized de novo on an Applied Biosystems 381A DNA synthesizer or equivalent type of machine.

The oligonucleotides may be labeled by any method known in the art. One preferred method is end-labeling, which can be performed as described in Maniatis et al., "Molecular Cloning: A Laboratory Manual", Cold Spring Harbor Laboratories, Cold Spring Harbor, New York (1982).

It is possible that in organisms having a relatively non-complex genome, only a

minimal complexity reduction step is necessary, and the genomic DNA may be directly analyzed or minimally reduced. This is particularly useful for screening tissue isolates to detect the presence of a bacterium or to identify the bacteria. Additionally, it is possible that, upon development of certain technical advances (e.g., more stringent hybridization, more sensitive detection equipment), even complex genomes may not need an extensive complexity reduction step.

Preferably, automated genotyping is performed. In general, genomic DNA of a well-characterized set of subjects, such as the CEPH families, is processed using PCR with appropriate primers to produce RCGs. The DNA is spotted onto one or more surfaces (e.g., multiple glass slides) for genotyping. This process can be performed using a microarray spotting apparatus which can spot more than 1,000 samples within a square centimeter area, or more than 10,000 samples on a typical microscope slide. Each slide is hybridized with a fluorescently tagged allele-specific SNP oligonucleotide under TMAC conditions analogous to those described below. The genotype of each individual can be determined by detecting the presence or absence of a signal for a selected set of SNP-ASOs. A schematic of the method is shown in Figure 4.

Once the complexity of genomic DNA obtained from an individual has been reduced, the resulting genomic DNA fragments can be attached to a solid support in order to be analyzed by hybridization. The RCG fragments may be attached to the slide by any method for attaching DNA to a surface. Methods for immobilizing nucleic acids have been described extensively, e.g., in US Patent Nos. 5,679,524; 5,610,287; 5,919,626; and 5,445,934. For instance, DNA fragments may be spotted onto poly-L-lysine- coated glass slides, and then crosslinked by UV irradiation. A second, more preferred method, which has been developed, involves including a 5′ amino group on each of the DNA fragments of the RCG. The DNA fragments are spotted onto silane-coated slides in the presence of NaOH in order to covalently attach the fragments to the slide. This method is advantageous because a covalent bond is formed between the fragments and the surface. Another method for accomplishing DNA fragment immobilization is to spot the RCG fragments onto a nylon membrane. Other methods of binding DNA to surfaces are possible and are well known to those of ordinary skill in the art. For instance, attachment to amino-alkyl-coated slides can be used. More detailed methods are described in the Examples below.

The surface to which the oligonucleotide arrays are conjugated is preferably a rigid or

semi-rigid support which may, optionally, have appropriate light absorbing or transmitting characteristics for use with commercially available detection equipment. Substrates which are commonly used and which have appropriate light absorbing or transmitting characteristics include, but are not limited to, glass, Si, Ge, GaAs, GaP, SiO_2, SiN_4, modified silicon, and polymers such as (poly)tetrafluoroethylene, (poly)vinylidenedifluoride, polystyrene, polycarbonate, or combinations thereof. Additionally, the surface of the support may be non-coated or coated with a variety of materials. Coatings include, but are not limited to, polymers, plastics, resins, polysaccharides, silica or silica-based materials, carbon, metals, inorganic glasses, and membranes.

In one embodiment. the SNP-ASOs are hybridized under standard hybridization conditions with RCGs covalently conjugated to a surface. Briefly, SNP-ASOs are labeled at their 5′ ends. A hybridization mixture containing the SNP-ASOs and, optionally, an isostabilizing agent, denaturing agent, or renaturation accelerant is brought into contact with an array of RCGs immobilized on the surface and the mixture and the surface are incubated under appropriate hybridization conditions. The SNP-ASOs which do not hybridize are removed by washing the array with a wash mixture (such as a hybridization buffer) to leave only hybridized SNP-ASOs attached to the surface. After washing, detection of the label (e.g., a fluorescent molecule) is performed. For example, an image of the surface can be captured (e.g., using a fluorescence microscope equipped with a CCD camera and automated stage capabilities, phosphoimager, etc.). The label may also, or instead, be detailed using a microarray scanner (e.g. one made by Genetic Microsystems). A microarray scanner provides image analysis which can be converted to a binary (i.e. +/-) signal for each sample using, for example, any of several available software applications (e.g., NIH image, ScanAnalyze, etc.) in a data format. The high signal/noise ratio for this analysis allows determination of data in this mode to be straightforward and easily automated. These data, once exported, can be manipulated to generate a format which can be directly analyzed by human genetics applications (such as CRI-MAP and LINKAGE via software). Additionally, the methods may utilize two or more fluorescent dyes which can be spectrally differentiated to reduce the number of samples to be analyzed. For instance, if four fluorescent dyes having spectral distinctions (e.g., ABI Prism dyes 6-FAM, HEX, NED, ROX) are used. Then four hybridization reactions can be carried out under a single hybridization condition. In other embodiments discussed in more detail below, the SNP-ASOs are conjugated to a surface and

hybridized with RCGs.

Conditions for optimal hybridization are described below in the Examples. In general, the SNP-ASO is present in a hybridization mixture at a concentration of from about 0.005 nanomoles per liter SNP-ASO hybridization mixture to about 50 nM SNP-ASO per ml hybridization mixture. More preferably, the concentration is from .5 nanomoles per liter to 1 nanomole per liter. A preferred concentration for radioactivity is 0.66 nanomoles per liter. The mixture preferably also includes a hybridization optimizing agent in order to improve signal discrimination between genomic sequences which are identically complementary to the SNP-ASO and those which contain a single mismatched nucleotide (as well as any neutral base etc. substitutions). Isostabilizing agents are compounds such as betaines and lower tetraalkyl ammonium salts which reduce the sequence dependence of DNA thermal melting transitions. These types of compounds also increase discrimination between matched and mismatched SNPs/genomes. A denaturing agent may also be included in the hybridization mixture. A denaturing agent is a composition that lowers the melting temperature of double stranded nucleic acid molecules, generally by reducing hydrogen bonding between bases or preventing hydration of nucleic acid molecules. Denaturing agents are well-known in the art and include, for example, DMSO, formaldehyde, glycerol, urea, formamide, and chaotropic salts. The hybridization conditions in general are those used commonly in the art, such as those described in Sambrook et al., "Molecular Cloning: A Laboratory Manual", (1989), 2nd Ed., Cold Spring Harbor, NY; Berger and Kimmel, "Guide to Molecular Cloning Techniques", *Methods in Enzymology*, (1987), Volume 152, Academic Press, Inc., San Diego, CA; and Young and Davis, (1983), *PNAS* (USA) 80:1194.

In general, incubation temperatures for hybridization of nucleic acids range from about 20°C to 75°C. For probes 17 nucleotides residues and longer, a preferred temperature range for hybridization is from about 50°C to 54°C. The hybridization temperature for longer probes is preferably from about 55°C to 65°C and for shorter probes is less than 52°C. Rehybridization may be performed in a variety of time frames. Preferably, hybridization of SNP and RCGs performed for at least 30 minutes.

Preferably, either or both of the SNP-ASO and the RCG are labeled. The label may be added directly to the SNP-ASO or the RCG during synthesis of the oligonucleotide or during generation of RCG fragments. For instance, a PCR reaction performed using labeled primers or labeled nucleotides will produce a labeled product. Labeled nucleotides (e.g., fluorescein-

labeled CTP) are commercially available. Methods for attaching labels to nucleic acids are well known to those of ordinary skill in the art and, in addition to the PCR method, include, for example, nick translation and end-labeling.

Labels suitable for use in the methods of the present invention include any type of label detectable by standard means, including spectroscopic, photochemical, biochemical, electrical, optical, or chemical methods. Preferred types of labels include fluorescent labels such as fluorescein. A fluorescent label is a compound comprising at least one fluorophore. Commercially available fluorescent labels include, for example, fluorescein phosphoramidides such as fluoreprime (Pharmacia, Piscataway, NJ), fluoredite (Millipore, Bedford, MA), FAM (ABI, Foster City, CA), rhodamine, polymethadine dye derivative, phosphores, Texas red, green fluorescent protein, CY3, and CY5. Polynucleotides can be labeled with one or more spectrally distinct fluorescent labels. "Spectrally distinct"fluorescent labels are labels which can be distinguished from one another based on one or more of their characteristic absorption spectra, emission spectra, fluorescent lifetimes, or the like. Spectrally distinct fluorescent labels have the advantage that they may be used in combination ("multiplexed"). Radionuclides such as ^{3}H, ^{125}I, ^{35}S, ^{14}C, or ^{32}P are also useful labels according to the methods of the invention. A plurality of radioactively distinguishable radionuclides can be used. Such radionuclides can be distinguished, for example, based on the type of radiation (e.g. α, β, or δ radiation) emitted by the radionuclides. The ^{32}P signal can be detected using a phosphoimager, which currently has a resolution of approximately 50 microns. Other known techniques, such as chemiluminescence or colormetric (enzymatic color reaction), can also be used.

By using spectrally distinct fluorescent probes , it is possible to analyze more than one locus a single hybridization mixture. The term "multiplexing" refers to the use of a set of distinct fluorescent labels in a single assay. Such fluorescent labels have been described extensively in the art, such as the fluorescent labels described in PCT Published Patent Application WO98/31834.

Fluorescent primers are a preferred method of labeling polynucleotides. The fluorescent tag is stable for more than a year. Radioactively labeled primers are stable for a shorter period. In addition, fluorescent primers may be used in combination if they are spectrally distinct, as discussed above. This allows multiple hybridizations to be detected in a single hybridization mixture. As a result, the total number of reactions needed for a genome-

wide scan is reduced. For example, for analysis of 1000 loci, 2000 hybridizations are needed (1000 loci x 2 polymorphisms/loci). The use of 4 fluorescently-labeled oligonucleotides will cut this number 4-fold and thus only 500 hybridizations will be needed.

In order to determine the genotype of an individual at a SNP locus, it is desirable to employ SNP allele-specific oligonucleotide hybridization. Preferably, two hybridization mixtures are prepared for each locus (or they can be performed together). The first hybridization mixture contains a labeled (e.g., radioactive or fluorescent) SNP-ASO (typically 17-21 nucleotide residues in length centered around the polymorphic residue). To increase specificity, a 20-50 fold excess of non-labeled oligonucleotides corresponding to another allele (referred to herein as a "complementary SNP-ASO") is included in the hybridization mixture. Use of the non-labeled complementary SNP-ASO can be avoided by using SNP-ASO containing a neutral base as described below. In the second hybridization mixture, the SNP-ASO that was labeled in the first mixture is not labeled, and the non-labeled SNP-ASO is labeled instead. Hybridization is performed in the presence of a hybridization buffer. The melting temperature of oligonucleotides can be determined empirically for each experiment. The pair of 2 oligonucleotides corresponding to different alleles of the same SNP (the SNP-ASOs and the complementary SNP-ASO) are referred to herein as a pair of allele-specific oligonucleotides (ASOs). Further experimental details regarding selecting and making SNP-ASOs are provided in the Examples section below.

In addition to the method described above, several other methods of allele specific hybridization may be used for hybridizing SNP-ASOs with RCGs. One method is to increase discrimination of SNPs in DNA hybridization by means of artificial mismatches. Artificial mismatches are inserted into oligonucleotide probes using a neutral base such as the base analog 3-nitropyrrole. A significant enhancement of discrimination is generally obtained, with a strong dependence of the enhancement on the spacing between mismatches.

In general, the methods described above are based on conjugation of genomic DNA fragments (i.e. a RCG) to a solid support. Hybridization analysis can also be performed with the SNP-ASO conjugated to the support (e.g. in an array). The oligonucleotide array is hybridized with one or more RCGs. Attaching of the SNP-ASOs or RCGs onto the support may be performed by any method known in the art. Many methods for attaching oligonucleotides to surfaces in arrays have been described, see, e.g. PCT Published Patent Application WO97/29212, US Patent Nos. 4,588,682; 5,667,976; and 5,760,130. Other

methods include, for example, using arrays of metal pins. Additionally, RCGs may be attached to the surface by the methods disclosed in the Examples below.

An "array" as used herein is a set of molecules arranged in a specific order with respect to a surface. Preferably the array is composed of polynucleotides (e.g. either SNP-ASOs or RCGs) attached to the surface. Oligonucleotide arrays can be used to screen nucleic acid samples for a target nucleic acid, which can be labeled with a detectable marker. A fluorescent signal resulting from hybridization between a target nucleic acid and a substrate-bound oligonucleotide provides information relating to the identity of the target nucleic acid by reference to the location of the oligonucleotide in the array on the substrate. Such a hybridization assay can generate thousands of signals which exhibit different signal strengths. These signals correspond to particular oligonucleotides of the array. Different signal strengths will arise based on the amount of labeled target nucleic acid hybridized with an oligonucleotide of the array. This amount, in turn, can be influenced by the proportion of AT-rich regions and GC-rich regions within the oligonucleotide (which determines thermal stability). The relative amounts of hybridized target nucleic acid can also be influenced by, for example, the number of different probes arrayed on the substrate, the length of the target nucleic acid, and the degree of hybridization between mismatched residues. Oligonucleotide arrays, in some embodiments, have a density of at least 500 features per square centimeter, but in practice can have much lower densities. A feature, as used herein, is an area of a substrate on which oligonucleotides having a single sequence are immobilized.

The oligonucleotide arrays of the invention may be produced by any method known in the art. Many such arrays are commercially available, and many methods have been described for producing them. One preferred method for producing arrays includes spatially directed oligonucleotide synthesis. Spatially directed oligonucleotide may be performed using light-directed oligonucleotide synthesis, microlithography, application by ink jet, microchannel deposition to specific location, and sequestration with physical barriers. Each of these methods is well-known in the art and has been described extensively. For instance, the light-directed oligonucleotide synthesis method has been disclosed in U.S. Patent Nos. 5,143,854; 5,489,678; and 5,571,639; and PCT applications having publication numbers WO90/15070; WO92/10092; and WO94/12305. This technique involves modification of the surface of the solid support with linkers and photolabile protecting groups using a photolithographic mask to produce reactive (e.g. hydroxyl) groups in the illuminated regions. A 3'-O-phosphoramide-

activated deoxynucleocide having a 5'-hydroxyl protected group is supplied to the surface such that coupling occurs at sites that were exposed to light. The substrate is rinsed, and the surface is illuminated with a second mask, and another activated deoxynucleotide is presented to the surface. The cycle is repeated until the desired set of products is obtained. After the cycle is finished, the nucleotides can be capped. Another method involves mechanically protecting portions of the surface and selectively deprotecting/coupling materials to the exposed portions of the surface, such as the method described in U.S. Patent No. 5,384,261. The mechanical means is generally referred to as a mask. Other methods for array preparation are described in PCT Published Patent Applications WO97/39151, WO98/20967, and WO98/10858, which describe an automated apparatus for the chemical synthesis of molecular arrays, U.S. Patent No. 5,143,854, Fodor et al., *Science* (1991), 251:767-777 and Kozal et al., *Nature Medicine*, v. 2, p. 753-759 (1996).

Hybridizing a SNP-ASO with an array of RCGs (or hybridizing a RCG with an array of SNP ASO) is followed by detection of hybridization. Part of the genotyping methods described herein is to determine if a positive or negative signal exists for each hybridization for an individual and then based on this information, determine the genotype for the corresponding SNP locus. This step is relatively straightforward, but varies depending on the method of detection. Essentially, all of the detection methods described here (fluorescent, radioactive, etc.) can be reduced to a digital image file, e.g. using a microarray reader or phosphoimager. Presently, there are several software products which will overlay a grid on an image and determine the signal strength value for each element of the grid. These values can be imported into a computer program, such as the Microsoft Corporation spreadsheet program designated Microsoft Excel™, with which simple analysis can be performed to assign each signal a manipulable value (e.g. 1 or 0 or + or -). Once this is accomplished, an individual's genotype can be described in terms of the pattern of hybridization of RCG fragments obtained from the individual with selected SNP ASO corresponding to disease-associated SNPs.

The array having labeled SNP-ASOs (or labeled RCGs) hybridized thereto can be analyzed using automated equipment. Automated equipment for analyzing arrays can include an excitation radiation source which emits radiation at a first wavelength, an optical detector, and a stage for securing the surface supporting the array. The excitation source emits excitation radiation which is focused on at least one area of the array and which induces

emission from fluorescent labels. The signal is preferably in the form of radiation having a different wavelength than the excitation radiation. Emitted radiation is collected by a detector, which generates a signal proportional to the amount of radiation sensed thereon. The array may then be moved so that a different area can be exposed to the radiation source to produce a signal. Once each area of the array has been scanned, a two-dimensional image of the array is obtained. Preferably, the movement of the array is accomplished using automated equipment, such as a multi-axis translation stage, such as one which moves the array at a constant velocity. In alternative embodiments, the array may remain stationary, and devices may be employed to cause scanning of the light over the stationary array.

One type of detection method includes a CCD imaging system, e.g. when the nucleic acids are labeled with fluorescent probes. Other detectors are well known to those of skill in the art and also, or alternatively, be used. CCD imaging systems for use with array detection have been described. For instance, a photodiode detector may be placed on the opposite side of the array from the excitation source. Alternatively, a CCD camera may be used in place of the photodiode detector to image the array. One advantage of using these systems is rapid read time. In general, an entire 50 x 50 centimeter array can be read in about 30 seconds or less using standard equipment. If more powerful equipment and efficient dyes are used, the read time may be reduced to less than 5 seconds.

Once the data is obtained, e.g. as a two-dimensional image, a computer can be used to transform the data into a displayed image which varies in color depending on the intensity of light emission at a particular location. Any type of commercial software which can perform this type of data analysis can be used. In general, the data analysis involves the steps of determining the intensity of the fluorescence emitted as a function of the position on the substrate, removing the outliers, and calculating the relative binding affinity. One or more of the presence, absence, and intensity of signal corresponding to a label is used to assess the presence or absence of an SNP corresponding to the label in the RCG. The presence and absence of one or more SNP's in a RCG can be used to assign a genotype to the individual. For example, the following depicts the genotype analysis of 3 individuals at a given locus at which an A/G polymorphism occurs:

Individual	SNP 1 Allele "A"	SNP 1 Allele "G"	Genotype
Larry	+	-	A/A*

Moe	-	+	G/G
Curly	+	+	A/G

As mentioned above, SNP analysis can be used to determine whether an individual has or will develop a particular phenotypic trait and whether the presence or absence of a specific allele correlates with a particular phenotypic trait. In order to determine which SNPs are related to a particular phenotypic trait, genomic samples are isolated from a group of individuals which exhibit the particular phenotypic trait, and the samples are analyzed for the presence of common SNPs. The genomic sample obtained from each individual is used to prepare a RCG. These RCGs are screened using panels of SNPs in a high throughput method of the invention to determine whether the presence or absence of a particular allele is associated with the phenotype. In some cases, it may be possible to predict the likelihood that a particular subject will exhibit the related phenotype. If a particular polymorphic allele is present in 30% of individuals who develop Alzheimer's disease, then an individual having that allele has a higher likelihood of developing Alzheimer's disease. The likelihood can also depend on several factors such as whether individuals not afflicted with Alzheimer's disease have this allele and whether other factors are associated with the development of Alzheimer's disease. This type of analysis can be useful for determining a probability that a particular phenotype will be exhibited. In order to increase the predictive ability of this type of analysis, multiple SNPs associated with a particular phenotype can be analyzed. Although values can be calculated, it is enough to identify that a difference exists.

It is also possible to identify SNPs which segregate with a particular disease. Multiple polymorphic sites may be detected and examined to identify a physical linkage between them or between a marker (SNP) and a phenotype. Both of these are useful for mapping a genetic locus linked to or associated with a phenotypic trait to a chromosomal position and thereby revealing one or more genes associated with the phenotypic trait. If two polymorphic sites segregate randomly, then they are either on separate chromosomes or are distant enough, with respect to one another on the same chromosome that they do not co-segregate. If two sites co-segregate with significant frequency, then they are linked to one another on the same chromosome. These types of linkage analyses are useful for developing genetic maps. See e.g., Lander et al., PNAS (USA) 83, 7353-7357 (1986), Lander et al., Genetics 121, 185-199 (1989). The invention is also useful for identifying polymorphic sites which do not segregate,

i.e., when one sibling has a chromosomal region that includes a polymorphic site and another sibling does not have that region.

Linkage analysis is often performed on family members which exhibit high rates of a particular phenotype or on patients suffering from a particular disease. Biological samples are isolated from each subject exhibiting a phenotypic trait, as well as from subjects which do not exhibit the phenotypic trait. These samples are each used to generate individual RCGs and the presence or absence of polymorphic markers is determined using panels of SNPs. The data can be analyzed to determine whether the various SNPs are associated with the phenotypic trait and whether or not any SNPs segregate with the phenotypic trait.

Methods for analyzing linkage data have been described in many references, including Thompson & Thompson, Genetics in Medicine (5th edition), W.B. Saunders Co., Philadelphia, 1991; and Strachan, "Mapping the Human Genome" in the Human Genome (Bios Scientific Publishers Ltd., Oxford) chapter 4, and summarized in PCT published patent application WO98/18967 by Affymetrix, Inc. Linkage analysis involving by calculating log of the odds values (LOD values) reveals the likelihood of linkage between a marker and a genetic locus at a recombination fraction, compared to the value when the marker and genetic locus are not linked. The recombination fraction indicates the likelihood that markers are linked. Computer programs and mathematical tables have been developed for calculating LOD scores of different recombination fraction values and determining the recombination fraction based on a particular LOD score, respectively. See e.g., Lathrop, PNAS, USA 81, 3443-3446 (1984); Smith et al., Mathematical Tables for Research Workers in Human Genetics (Churchill, London, 1961); Smith, Ann. Hum. Genet. 32, 127-1500 (1968). Use of LOD values for genetic mapping of phenotypic traits is described in PCT published patent application WO98/18967 by Affymetrix, Inc. In general, a positive LOD score value indicates that two genetic loci are linked and a LOD score of +3 or greater is strong evidence that two loci are linked. A negative value suggests that the linkage is less likely.

The methods of the invention are also useful for assessing loss of heterozygosity in a tumor. Loss of heterozygosity in a tumor is useful for determining the status of the tumor, such as whether the tumor is an aggressive, metastatic tumor. The method is generally performed by isolating genomic DNA from tumor sample obtained from a plurality of subjects having tumors of the same type, as well as from normal (i.e., non-cancerous) tissue obtained from the same subjects. These genomic DNA samples are used to generate RCGs

which can be hybridized with a SNP-ASO, for example using the surface array technology described herein. The absence of a SNP allele in the RCG generated from the tumor compared to the RCG generated from normal tissue indicates whether loss of heterozygosity has occurred. If a SNP allele is associated with a metastatic state of a cancer, the absence of the SNP allele can be compared to its presence or absence in a non-metastatic tumor sample or a normal tissue sample. A database of SNPs which occur in normal and tumor tissues can be generated and an occurrence of SNPs in a patient's sample can be compared with the database for diagnostic or prognostic purposes.

It is useful to be able to differentiate non-metastatic primary tumors from metastatic tumors, because metastasis is a major cause of treatment failure in cancer patients. If metastasis can be detected early, it can be treated aggressively in order to slow the progression of the disease. Metastasis is a complex process involving detachment of cells from a primary tumor, movement of the cells through the circulation, and eventual colonization of tumor cells at local or distant tissue sites. Additionally, it is desirable to be able to detect a predisposition for development of a particular cancer such that monitoring and early treatment may be initiated. Many cancers and tumors are associated with genetic alterations. For instance, an extensive cytogenetic analysis of hematologic malignancies such as lymphomas and leukemias have been described, see e.g., Solomon et al., Science 254, 1153-1160, 1991. Many solid tumors have complex genetic abnormalities requiring more complex analysis.

Solid tumors progress from tumorigenesis through a metastatic stage and into a stage at which several genetic aberrations can occur. e.g., Smith et al., Breast Cancer Res. Terat., 18 Suppl. 1, S5-14, 1991. Genetic aberrations are believed to alter the tumor such that it can progress to the next stage, i.e., by conferring proliferative advantages, the ability to develop drug resistance or enhanced angiogenesis, proteolysis, or metastatic capacity. These genetic aberrations are referred to as "loss of heterozygosity." Loss of heterozygosity can be caused by a deletion or recombination resulting in a genetic mutation which plays a role in tumor progression. Loss of heterozygosity for tumor suppressor genes is believed to play a role in tumor progression. For instance, it is believed that mutations in the retinoblastoma tumor suppressor gene located in chromosome 13q14 causes progression of retinoblastomas, osteosarcomas, small cell lung cancer, and breast cancer. Likewise, the short arm of chromosome 3 has been shown to be associated with cancer such as small cell lung cancer, renal cancer and ovarian cancers. For instance, ulcerative colitis is a disease which is

associated with increased risk of cancer presumably involving a multistep progression involving accumulated genetic changes (US Patent No. 5,814,444). It has been shown that patients afflicted with long duration ulcerative colitis exhibit an increased risk of cancer, and that one early marker is loss of heterozygosity of a region of the distal short arm of chromosome 8. This region is the site of a putative tumor suppressor gene that may also be implicated in prostate and breast cancer. Loss of heterozygosity can easily be detected by performing the methods of the invention routinely on patients afflicted with ulcerative colitis. Similar analyses can be performed using samples obtained from other tumors known or believed to be associated with loss of heterozygosity.

The methods of the invention are particularly advantageous for studying loss of heterozygosity because thousands of tumor samples can be screened at one time. Additionally, the methods can be used to identify new regions of loss that have not previously been identified in tumors.

The methods of the invention are useful for generating a genomic pattern for an individual genome of a subject. The genomic pattern of a genome indicates the presence or absence of polymorphisms, for example, SNPs, within a genome. Genomic DNA is unique to each individual subject (except identical twins). Accordingly, the more polymorphisms that are analyzed for a given genome of a subject, the higher probability of generating a unique genomic pattern for the individual from which the sample was isolated. The genomic pattern can be used for a variety of purposes, such as for identification with respect to forensic analysis or population identification, or paternity or maternity testing. The genomic pattern may also be used for classification purposes as well as to identify patterns of polymorphisms within different populations of subjects.

Genomic patterns may be used for many purposes, including forensic analysis and paternity or maternity testing. The use of genomic information for forensic analysis has been described in many references, see e.g., National Research Council, The Evaluation of Forensic DNA Evidence (EDS Pollard et al., National Academy Press, DC, 1996). Forensic analysis of DNA is based on determination of the presence or absence of alleles of polymorphic regions within a genomic sample. The more polymorphisms that are analyzed, the higher probability of identifying the correct individual from which the sample was isolated.

In an embodiment of the invention, when a biological sample, such as blood or sperm, is found at a crime scene, DNA can be isolated and RCGs can be prepared. This RCG can

then be screened with a panel of SNPs to generate a genomic pattern. The genomic pattern can be matched with a genomic pattern produced from a suspect or compared to a database of genomic patterns which has been compiled. Preferably, the SNPs used in the analysis are those in which the frequency of the polymorphic variation (allelic frequency) has been determined, such that a statistical analysis can be used to determine the probability that the sample genome matches the suspect's genome or a genome within the database. The probability that two individuals have the same polymorphic or allelic form at a given genetic site is described in detail in PCT published patent application WO98/18967, the entire contents of which are hereby incorporated by reference. Briefly, this probability defined as P(ID) can be determined by the equation:

$$P(ID)=(x^2)^2+(2xy)^2+(y^2)^2$$

x and y in the equation represent the frequency that an allele A or B will occur in a haploid genome.

The calculation can be extended for more polymorphic forms at a given locus. The predictability increases with the number of polymorphic forms tested. In a locus of n alleles, a binomial expansion is used to calculate P(ID). The probabilities of each locus can be multiplied to provide the cumulative probability of identity and from this the cumulative probability of non-identity for a particular number of loci can be calculated. This value indicates the likelihood that random individuals have the same loci. The same type of quantitative analysis can be used to determine whether a subject is a parent of a particular child. This type of information is useful in paternity testing, animal breeding studies, and identification of babies or children whose identity has been confused, e.g., through adoption or inadequate record keeping in a hospital, or through separation of families by occurrences such as earthquake or war.

The genomic pattern may be used to generate a genomic classification code (GNC). The GNC may be represented by one or more data signals and stored as part of a data structure on a computer-readable medium, for example, a database. The stored GNCs may be used to characterize, classify, or identify the subjects for which the GNCs were generated. Each GNC may be generated by representing the presence or absence of each polymorphism with a computer-readable signal. These signals may then be encoded, for example, by performing a function on the signals.

Accordingly, the GNCs may be used as part of a classification or identification system

for subjects such as, for example, humans, plants, or animals. As discussed above, the more polymorphisms that are analyzed for a given genome of a subject, the higher probability of generating a unique genomic pattern for the individual from which the sample was isolated, and consequently, the higher the probability that the GNC uniquely identifies an individual. In such a system, a data structure may include a plurality of entries, for example, data records or table entries, where each entry identifies an individual. Each entry may include the GNC generated for the individual as well as other. The GNC or portions thereof may then be stored in an index data structure, for example, another table. A portion of a GNC may be indexed so that each GNC may be further classified by a portion of its genomic pattern as opposed to only the entire genomic pattern.

The data structures may then be searched to identify an individual who has committed a crime. For example, if a biological sample from the individual (such as blood) is recovered from the crime scene, the GNC of the individual may generated by the methods described herein, and a database of records including GNCs searched until a match is found. Thus, the GNCs may be used to classify individuals within a group such as soldiers in the armed forces, cattle in a herd, or produce within a specific crop. For example, the armed forces may generate a database containing the GNC of each soldier, and the database could be used to identify the soldier if necessary. Likewise, a database could be generated where records and indexes of the database include the GNCs of individual animals within a herd of cattle, so that lost or stolen animals could later be identified and returned to the proper owner.

The code may optionally be converted into a bar code or other human- or machine-readable form. For example, each line of a bar code may indicate the presence of specific polymorphisms or groups of specific polymorphisms for a particular subject.

Additionally, it is useful to be able to identify the genus, species, or other taxonomic classification to which an organism belongs. The methods of the invention can accomplish this in a high throughput manner. Taxonomic identification is useful for determining the presence and identity of a pathogenic organism such as a virus, bacteria, protozoa, or multicellular parasites in a tissue sample. In most hospitals, bacteria and other pathogenic organisms are identified based on morphology, determination of nutritional requirements or fermentation patterns, determination of antibiotic resistance, comparison of isoenzyme patterns, or determination of sensitivity to bacteriophage strains. These types of methods generally require approximately 48 to 72 hours to identify the pathogenic organism. More

recently, methods for identifying pathogenic organisms have been focused on genotype analysis, for instance, using RFLPs. RFLP analysis has been performed using hybridization methods (such as southern blots) and PCR assays.

The information generated according to the methods of the invention and in particular the GNCs, can be included in a data structure, for example, a database, on computer-readable medium, wherein the information is correlated with other information pertaining to the genomes or the subjects or types of subjects, from which the genomes are obtained. FIG. 5 shows a computer system 100 for storing and manipulating genomic information. The computer system 100 includes a genomic database 102 which includes a plurality of records 104a-*n* storing information corresponding to a plurality of genomes. Each of the records 104a-*n* may store genetic information about each genome or an RCG generated therefrom. The genomes for which information is stored in the genomic database 102 may be any kind of genomes from any type of subject. For example, the genomes may represent distinct genomes of individual members of a species, particular classes of the individuals, ie., army, prisoners, etc.

An example of the format of a record 200 in the genomic database 102 (i.e., one of the records 104a-*n*) is shown in FIG. 6A. As shown in FIG. 6A, the record 200 includes a genome identifier (Genome ID) 202 that identifies the genome corresponding to the record 200. If enough polymorphisms of the genome were analyzed to generate the spectral pattern (such that the possibility that the GNC uniquely identifies the genome is high), or if a group to which the genome belongs has few enough members, than the GNC of the genome could serve as the Genome ID 202. The record 202 also may include genomic information fields 204a-*n*. The genomic information may be any information associated with the genome identified by the Genome ID 202 such as, for example, a GNC, a portion of a GNC, the presence or absence of a particular SNP, a genetic attribute (genotype), a physical attribute (phenotype), a name, a taxonomic identifier, a classification of the genome, a description of the individual from which the genome was taken, a disease of the individual, a mutation, a color, etc. Each information field 204a-n may be used as an entry in an index data structure that has a structure similar to record 200. For example, each entry of the index data structure may include an indexed information field as a first data element, and one or more Genome IDs 202 as additional elements, such that all elements that share a common attribute are stored in a common data structure. The format of the record 200 shown in FIG. 6A is merely an

example of a format that may be used to represent genomes in the genomic database 102. The amount of information stored for each record 200, the number of records 200, and the number of fields indexed may vary.

Further, each information field 204a-n may include one or more fields itself, and each of these fields themselves may include more fields, etc. Referring to FIG. 6B, an embodiment of the information field 204a is shown. The information field 204a includes a plurality of fields 206a-*m* for storing more information about the information represented by information field 204a. Although the following description refers to the fields 206a-*m* of the gene ID 204a, such description is equally applicable to information fields 204b-*n*. For example, if information field 204a represented a GNC of the genome corresponding to the genome ID 202, then each of the fields 206a-m may represent a portion of the GNC, a particular SNP of the genomic pattern from which the GNC was generated, a group of such SNPs, a description of the GNC, a description of a one of the SNPs, etc.

The fields 206a-*m* of the gene ID 204a may store any kind of value that is capable of being stored in a computer readable medium such as, for example, a binary value, a hexadecimal value, an integral decimal value, or a floating point value.

A user may perform a query on the genomic database 102 to search for genomic information of interest, for example, all genomes having a GNC that matches the GNC of a murder suspect. In another example, it may be known that a biological sample contains a particular sequence. That sequence can be compared with sequences in the database to identify information such as which individual the sample was isolated from, or whether the genetic sequence corresponds to a particular phenotypic trait. For example, the user may search the genomic database 102 for genetic matches to identify an individual, genotypes which correlate with a particular phenotype, genotypes associated with various classes of individuals etc. Referring to FIG. 5, a user may provide user input 106 indicating genomic information for which to search to a query user interface 108. The user input 106 may, for example, indicate an SNP for which to search using a standard character-based notation. The query user interface 108 may, for example, provide a graphical user interface (GUI) which allows the user to select from a list of types of accessible genomic information using an input device such as a keyboard or a mouse.

The query user interface 108 generates a search query 110 based on the user input 106. A search engine 112 receives the search query 110 and generates a mask 114 based on the

search query. Example formats of the mask 114 and ways in which the mask 114 may be used to determine whether the genomic information specified by the mask 114 matches genomic information of genomes in the genomic database 102 are described in more detail below with respect to FIG. 7. The search engine 112 determines whether the genomic information specified by the mask 114 matches genomic information of genomes stored in the genomic database 102. As a result of the search, the search engine 112 generates search results 116 indicating whether the genomic database 102 includes genomes having the genomic information specified by the mask 114. The search results 116 may also indicate which genomes in the genomic database 102 have the genomic information specified by the mask 114.

If, for example, the user input 106 specified a sequence of a gene, a GNC, or an SNP, the search results 116 may indicate which genomes in the genomic database 102 include the specified sequence, GNC, or SNP. If the user input 106 specified particular genetic information concerning a genome (e.g., enough to identify an individual), the search results 116 may indicate which individual genome listed in the genomic database 102 matches the particular information, thus identifying the individual from whom the sample was taken. Similarly, if the user input 106 specified genetic sequences which are not adequate to specifically identify the individual, the search results 116 may still be adequate to identify a class of individuals that have genomes in the genomic database 102 that match the genetic sequence. For example, the search results may indicate that the genomic information of genomes of all caucasian males matches the specified genetic sequence.

FIG. 7 illustrates a process 300 that may be used by the search engine 112 to generate the search results 116. The search engine 112 receives the search query 110 from the query user interface 108 (step 302). The search engine 112 generates the mask 114 generated based on the search query 110 (step 304). The search engine 112 performs a binary operation on one or more of the records 104a-*n* in the genomic database 102 using the mask 114 (step 306). The search engine 112 generates the search results 116 based on the results of the binary operation performed in step 306 (step 308).

A computer system for implementing the system 100 of FIG. 5 as a computer program typically includes a main unit connected to both an output device which displays information to a user and an input device which receives input from a user. The main unit generally includes a processor connected to a memory system via an interconnection mechanism. The

input device and output device also are connected to the processor and memory system via the interconnection mechanism.

One or more output devices may be connected to the computer system. Example output devices include a cathode ray tube (CRT) display, liquid crystal displays (LCD), printers, communication devices such as a modem, and audio output. One or more input devices may be connected to the computer system. Example input devices include a keyboard, keypad, track ball, mouse, pen and tablet communication device, and data input devices such as sensors. The invention is not limited to the particular input or output devices used in combination with the computer system or to those described herein.

The computer system may be a general purpose computer system which is programmable using a computer programming language, such as for example, C++, Java, or other language, such as a scripting language or assembly language. The computer system may also include specially programmed, special purpose hardware such as, for example, an application-specific integrated circuit (ASIC). In a general purpose computer system, the processor is typically a commercially available processor, of which the series x86, Celeron, and Pentium processors, available from Intel, and similar devices from AMD and Cyrix, the 680X0 series microprocessors available from Motorola, the PowerPC microprocessor from IBM and the Alpha-series processors from Digital Equipment Corporation, are examples. Many other processors are available. Such a microprocessor executes a program called an operating system, of which Windows NT, Linux, UNIX, DOS, VMS and OS8 are examples, which controls the execution of other computer programs and provides scheduling, debugging, input/output control, accounting, compilation, storage assignment, data management and memory management, and communication control and related services. The processor and operating system define a computer platform for which application programs in high-level programming languages are written.

A memory system typically includes a computer readable and writeable nonvolatile recording medium, of which a magnetic disk, a flash memory, and tape are examples. The disk may be removable such as, for example, a floppy disk or a read/write CD, or permanent, known as a hard drive. A disk has a number of tracks in which signals are stored, typically in binary form, i.e., a form interpreted as a sequence of one and zeros. Such signals may define an application program to be executed by the microprocessor, or information stored on the disk to be processed by the application program. Typically, in operation, the processor causes

data to be read from the nonvolatile recording medium into an integrated circuit memory element, which is typically a volatile, random access memory such as a dynamic random access memory (DRAM) or static memory (SRAM). The integrated circuit memory element allows for faster access to the information by the processor than does the disk. The processor generally manipulates the data within the integrated circuit memory and then copies the data to the disk after processing is completed. A variety of mechanisms are known for managing data movement between the disk and the integrated circuit memory element, and the invention is not limited to any particular mechanism. It should also be understood that the invention is not limited to a particular memory system.

The invention is not limited to a particular computer platform, particular processor, or particular high-level programming language. Additionally, the computer system may be a multiprocessor computer system or may include multiple computers connected over a computer network. It should be understood that each module (e.g. 108, 112) in FIG. 5 may be a separate module of a computer program, or may be a separate computer program. Such modules may be operable on separate computers. Data (e.g. 102, 106, 110, 114, and 116) may be stored in a memory system or transmitted between computer systems. The invention is not limited to any particular implementation using software, hardware, firmware, or any combination thereof. The various elements of the system, either individually or in combination, may be implemented as a computer program product tangibly embodied in a machine-readable storage device for execution by a computer processor. Various steps of the process, for example, steps 302, 304, 306, and 308 of FIG. 7, may be performed by a computer processor executing a program tangibly embodied on a computer-readable medium to perform functions by operating on input and generating output. Computer programming languages suitable for implementing such a system include procedural programming languages, object-oriented programming languages, and combinations of the two.

The invention also encompasses compositions. One composition of the invention is a plurality of RCGs immobilized on a surface, where the plurality of RCGs are prepared by DOP-PCR. Another composition is a panel of SNP-ASOs immobilized on a surface, wherein the SNPs are identified by using RCGs as described above.

The invention also includes kits having a container housing a set of PCR primers for reducing the complexity of a genome and a container housing a set of SNP-ASOs,

The invention also encompasses compositions. One composition of the invention is a plurality of RCGs immobilized on a surface, where the plurality of RCGs are prepared by DOP-PCR. Another composition is a panel of SNP-ASOs immobilized on a surface, wherein the SNPs are identified by using RCGs as described above.

The invention also includes kits having a container housing a set of PCR primers for reducing the complexity of a genome and a container housing a set of SNP-ASOs, particularly wherein the SNPs are present with a frequency of at least 50 or 55% in a RCG made using the primer set. In some kits, the set of PCR primers are primers for DOP-PCR and preferably the DOP-PCR primer has the tag-$(N)_x$-TARGET structure described herein, i.e., wherein the TARGET includes at least 7 arbitrarily selected nucleotide residues, wherein x is an integer from 3 to 9, and wherein each N is any nucleotide residue and wherein tag is a polynucleotide as described above. In some embodiments the SNPs in the kit are attached to a surface such as a slide.

SNPs identified according to the methods of the invention using the B1 5' rev primer include the following:

B1 5'rev ATTAAAGGCGTGCGCCACCATGCC (SEQID #13)

locus	ASO	Allele	Strain	(SEQID #)
1	tttatgAaggCataaaaa	A	129/	14
	tttatgGaggCataaaaa	B	B6-DBA	15
	tttatgAaggTataaaaa	C	Spre	16
2	ctgggctgTattcattt	A	129-DBA	17
	ctgggctgCattcattt	B	B6	18
	tctGcctccTGagtgct	C	B6-129-DBA	19
	tctAcctccCAagtgct	D	Spre	20
3	tagctagaAtcaagctt	A	B6	21
	tagctagaGtcaagctt	B	DBA-Spre	22
4	gctgtgcAACaaatcac	A	129/	23
	cagctgtgc---aaatcacc	B	B6	24
5	tttcgtga-tgtttctat	A	129-Spre	25
	tttcgtgaAtgtttcta	B	B6-DBA	26

6	cactgtctAcatcttta	A	B6-129	27
	cactgtctCcatcttta	B	DBA-Spre	28
7	taacattcTtgaagcca	A	129-DBA-Spre	29
	taacattcCtgaagcca	B	B6	30
8	gcttccaTttcctaagg	A	129-DBA	31
	gcttccaCttcctaagg	B	B6	32
9	aggaatgGcAataatcc	A	B6-129	33
	aggaatgGcGataatcc	B	DBA	34
	aggaatgAcAataatcc	C	Spre	35
	ttaaattcGtaaatgga	D	B6-129-DBA	36
	ttaaattcAtaaatgga	E	Spre	37
10	taacattcTtgaagcca	A	129-DBA-Spre	38
	taacattcCtgaagcca	B	B6	39
11	ttcTGtgActccaCttg	A	129	40
	ttcTGtgActccaTttg	B	B6-DBA	41
	ttcCCtgTctccaTttg	C	Spre	42
12	gtagtttgCcaggaacc	A	129-Spre	43
	gtagtttgTcaggaacc	B	B6-DBA	44
13	tgctactcctctctactcg	A	129	45

	tgctattcctctctgctcg	B	B6-DBA-Spre	46
	cttgatcaccctctgatga	C	B6-129-DBA	47
	cttggtcaccctctaatga	D	Spre	48
14	gaggtggtgcagagtga	A	129-DBA	49
	gaggtggcgcagagtga	B	B6	50
	gaggtggcccagagtga	C	Spre	51
15	cccactgaaccgcacag	A	129-DBA	52
	cccactgagctgcacag	B	B6	53
	cccactcagccgcacag	C	Spre	54
16	tgaagacacagccagcc	A	129-DBA	55
	tgaagacgcagccagcc	B	B6	56
	tgaagacgaagccagcc	C	Spre	57
17	agaagttggtaccaggg	A	129/FVB/F1/cast/spre	58
	agaagttgttaccaggg	B	B6	59
18	tatgattacgtaatgtt	A	129/B6/F1	60
	tatgattatgtaatgtt	B	FVB/F1	61
19	atgattccagtgagtta	A	129/B6	62
	atgattcctgtgagtta	B	FVB/F1	63
	catactattaacactggaa	C	Cast-129	64
	catattattaacacaggaa	D	Spre	65
20	gtcaagaacaggcaata	A	129/b6/f1/FVB	66
	gtcaagaataggcaata	B	f1	67
	cagactagggaaccttc	C	129	68
	cagacgagggaaccttc	E	Spre	69
	cagactagggagccttc	D	Cast	70
21	tgtccagttgtttgcat	A	129/	71
	tgtccagtcgtttgcat	B	b6/fvb/f1	72
	ggggtagccagtttggt	C	Cast-129	73
	ggggtagcaagtttggt	D	Spre	74
22	caggaagctgtagctcc	A	129/f1	75
	caggaagccgtagctcc	B	b6/fvb	76
	cctgagcctgtctacct	C	Cast-129	77
	cctgagcccgtctacct	D	Spre	78
23	taacattcttgaagcca	A	129/FVB/F1/cast/spre	79
	taacattcctgaagcca	B	B6	80

24	ccaactgaaccgcacag	A	129/FVB	81
	ccaactgagctgcacag	B	B6	82
	gagctagctcacacattct	C	Cast-129	83
	gagttagctcacacgttct	D	Spre	84
25	acgggggggtggcgtta	A	129/f1	85
	acgggggg-tggcgttaa	B	b6/fvb/cast/spre	86
	tagacagccagcgcgtcac	C	Cast-129	87
	tagatagccagcgcatcac	D	Spre	88
26	gcttttcttgagagtggc	A	129/b6	89
	gcttttctttagagtggc	B	fvb	90
	gcttttcgtgagagtggc	C	f1	91
27	ctacagataaagttata	A	129/b6/fvb/f1	92
	ctacagatgaagttata	B	f1	93
	tagacctgctgctatct	C	Cast-129	94
	tagacctgttgctatct	D	Spre	95
28	tgttgttctggcctcca	A	129/F1	96
	tgttgttttggcctcca	B	B6	97
	ttctgagaatttgttag	C	129/B6	98
	ttctgagagtttgttag	D	F1/spre	99
29	caggaagcagtagctcc	A	129	100
	caggaagccgtagctcc	B	B6/FVB/F1	101
	agagtcaggtaagttgc	C	Cast-129	102
	agagtcagataagttgc	D	Spre	103
30	agatttcaaaaagtttt	A	129/b6	104
	agattccaaaaggtttt	B	f1	105
	agatttcaaaaagtttt	C	fvb	106
	cctgaggggagcaatca	D	Cast-129	107
	cctgagggaagcaatca	E	Spre	108
31	aaggtaagataactaag	A	129.f1	109
	aaggtaaggtaactaag	B	b6/fvbn	110
	ggactacacagagaaac	C	Cast-129	111
	ggactacatagagaaac	D	Spre	112
32	cccaggctacacgaggg	A	129/fvb/f1	113
	cccaggctacatgaggg	B	b6	114
	cttaccagttgtgagac	C	129	115

	cttaccacttgtgagac	D	Spre	116
	cttaccagtcgtgagac	E	Cast	117
33	ctgccctcaggtcttta	A	129	118
	ctgccctccggtcttta	B	b6/fvbn	119
	gcaataaaattgtttta	C	Cast-129	120
	gcaatgagatcgtttta	D	Spre	121
34	tgttctgtggagacccc	A	129/fvbn/f1/cast/spre	122
	tgttctgtagagacccc	B	b6	123
35	cacattgaatcaaagcc	A	129/b6/fvbn/f1	124
	cacattgagtcaaagcc	B	f1	125
	ggactacccacccgttc	C	129	126
	gcgactgc--acccattct	E	Spre	127
	gcgactgccccc--attct	D	Cast	128
36	cctgggccagccaggaa	A	129/b6/cast	129
	cctgggcctgccaggaa	B	fvbn/f1/spre	136
37	ccccaggtaaccatctt	A	129/f1	131
	ccccaggtgaccatctt	B	b6/fvbn/cast/spre	132
	ttctgtatattagctga	C	Cast-129	133
	tttctatattaa--ctgac	D	Spre	134
38	ggacccggacggtcttc	A	129/b6	135
	ggacccggtcggtcttc	B	bvb/f1	136
	gtccctaatgttagcat	C	Cast-129	137
	gtccccaatgtcagcat	D	Spre	138
39	acgggggggtggcgtta	A	129/f1	139
	acgggggg-tggcgttaa	B	b6/fvbn/cast/spre	140
	tagacagccagcgcgtcac	C	Cast	141
	tagatagccagcgcatcac	D	Spre	142
40	gattcttcgtgttcctt	A	129-b6-F1	143
	gattcttcatgttcctt	B	FVBN-Cast-Spre	144
41	tgtaaaaacttagaata	A	129/b6/f1	145
	tgtaaaaatttagaata	B	fvbn/cast/spre	146
42	tgtgaaagcgctcccaa	A	129/fvbn/f1/cast/spre	147
	tgtgaaagtgctcccaa	B	b6	148
43	caaaggctcagagaatc	A	129/b6/f1	149
	caaaggcttagagaatc	B	fvbn	150

	ttaattctctccaaaca	C	129/b6/fvb/f1	151
	ttaaggctctccggaca	D	f1	152
44	ctgccaccgtgcacaca	A	129/b6	153
	ctgccaccatgcacaca	B	fvbn/f1	154
	ccaaatattctgattcc	C	129-Spre	155
	ccaaatattctttttt	D	Cast	156
45	atgagctgaccctccct	A	129/B6/F1	157
	atgagctgcccctccct	B	FVB	158
	acactaggtaaaagctc	C	129/B6/FVB/F1	159
	acactaggcaaaagctc	D	F1	160
	agacaccacgaccgagg	E	129-Spre	161
	agacaccaagaccgagg	F	Cast	162
46	gcagcgtccggttaagt	A	129/f1	163
	gcagcgtctggttaagt	B	b6/fvbn/f1	164
	cagatactacaaggatg	C	129	165
	tacagatac---aaggatgc	D	SPRE/Cast	166
47	tcagctagtgtatctgt	A	129/FVB/F1	167
	tcacctagtgtatttgt	B	B6/F1	168
	ttttttatttttggatt	C	129-Cast	169
	tttt-aatttttggattt	D	Spre	170
48	gatattgttttcattta	A	129/	171
	gatattgtcttcattta	B	b6/fvbn/f1	172
49	agacccggtgctggtgt	A	129/b6	173
	agacccggcgctggtgt	B	fvbn/f1/cast	174
50	cttctaagctttgtctt	A	129/fvb/f1/cast/spre	175
	cttctaagttttgtctt	B	b6/f1	176
51	agttggcaaccagcatg	A	129/	177
	agttggcatccagcatg	B	b6/fvbn/f1	178
	ggtgaaatggtaattac	C	129-Cast	179
	ggtgaaatagtaattac	D	Spre	180
52	acgggatataacgagtt	A	129/FVB/F1	181
	acgggatacaacgagtt	B	B6/cast/spre	182

	gggatacaacgagtttc	C	129-Cast	183
	gggatacaccgagtttc	D	Spre	184
53	gtatcttgggtgtcctg	A	129/FVB/F1	185
	gtaacttgggtgttctg	B	B6/F1/spre	186
	gggtgtcctgccccatc	C	129	187
	gggtgttctgttttatc	D	Spre	188
54	tgtccagttgttttgca	A	129	189
	tgtccagtcgttttgca	B	B6/FVB/F1/spre	196
	aagacagccggaactct	C	129...	191
	aagacagcaggaactct	D	Spre	192
55	tgataggaccaaagaga	A	129/b6/f1	193
	cgataggactaaagaga	B	fvbn/f1	194
	tccaaagccagggccca	C	129	195
	tccaaattcagggccca	D	Spre	196
56	cctgggccagccagaag	A	129/B6/cast	197
	cctgggcctgccagaag	B	FVB/F1/spre	198
57	gattctctgagcctttg	A	129/b6/f1	199
	gattctctaagcctttg	B	fvbn	200
	taccatttttagatga	C	129...	201
	taccatttcttagatga	D	Spre	202
58	ctggaagggcagtgaat	A	129	203
	tctgga-cgagggtgaat	B	B6/FVB	204
59	tagttgcagcacaaatg	A	129/B6	205
	tagttgtagcacaaatg	B	FVB/F1	206
60	acactaccgcacagagc	A	129/b6/fvbn/f1	207
	acactaccacacagagc	B	f1	208
	aataataagtaaataag	C	129/	209
	aataataaataaataag	D	cast	210
61	tggcagtagttgttcat	A	129/b6	211
	tggcagtaattgttcat	B	fvbn/f1	212
	aggtatgacgtcataag	C	129-Cast	213
	aggtatgatgtcataag	D	Spre	214
62	gttgttgttgaagattt	A	129/fvbn/f1	215
	ttgttgttg---aagattta	B	b6/f1	216

	gatagtacaggtgttgtca	C	129...	217
	gatggtacaggtgtcgtca	D	Spre	218
63	aatataatgtaacagga	A	129/F1	219
	aatataatataacagga	B	B6/FVB/F1	220
64	ttaaccatttatctgat	A	129/FVB	221
	ttaaccatatatctgat	B	B6/F1	222
65	agagcccagcaaagttc	A	129/B6	223
	agagcccaacaaagttc	B	FVB/F1	224
	atcccgaaccggggaaaat	C	129-b6	225
	atcccaaaccgggggaaat	D	cast-spre	226
66	atgacaccaccacaacc	A	129	227
	atgacaccgccacaacc	B	B6/FVB/F1	228
67	aggcaaacagatataac	A	129/FVB/F1	229
	aggcaaacggatataac	B	B6/cast/spre	230
	tgtattcactaataaga	C	129-Cast	231
	tgtattcattaataaga	D	Spre	232
68	ttggcgtatacttcata	A	129/B6/F1	233
	ttggcgtacacttcata	B	FVB	234
	ctcaccacgctccatct	C	129	235
	ctcaccaccctccatct	D	Cast-Spre	236
69	atatctaaa----ggcacag	A	129/FVB	237
	tatctacataaaggcac	B	B6/F1/cast/spre	238
	gtgtctcctagtctccc	C	B6-Cast	239
	gtgtctcccagtctccc	D	Spre	240
70	atgagctgaccctccct	A	129/B6/F1	241
	atgagctgcccctccct	B	FVB/F1	242
	ggacaacatttaattgg	C	129-Cast	243
	ggacaacacttaattgg	D	Spre	244
71	gctttaaaatttttatt	A	129	245
	gctttaaatttttttatt	B	B6/FVB/F1	246
	aaatttgttcctaaatg	C	129	247
	aaatttgtacctaaatg	D	Cast-Spre	248
72	gtgttgttctggcctcc	A	129/FVB/spre	249
	gtgttgttttggcctcc	B	B6/F1	250

73	tgaatgacaaaaagaca	A	129/B6/FVB	251
	tgaatgacgaaaagaca	B	F1/cast	252
B2 5'Rev	ACTGAGCCATCTCWCCAG W=A+T			
101	acttaacttaagctggc	A	129/	253
	gtacttaa-----gctggcctg	B	b6/fvb/f1	254
102	actctaatatcccacag	A	129/fvbn/f1	255
	actctaatctcccacag	B	b6	256
	cggatcggctctagttc	C	129/cast	257
	cggatcagctctagttc	D	spre	258
103	tcaaaccaataaggagg	A	129/b6/fvb/f1	259
	tcaaaccagtaaggagg	B	f1	260
104	gtgtgtgtgtgggggggg	A	129/f1	261
	gtgtgtgtg---ggggggggt	B	b6/fvbn	262
	cttaataataatttcat	C	129/cast	263
	cttaataacaatttcat	D	spre	264
105	gtgtctccatatgtgtg	A	129/b6/f1	265
	gtgtctacacatgtgtg	B	fvbn	266
106	aactcatcatgatggtt	A	129/	267
	aactcataatgatggtt	B	b6/fvbn/f1	268
	aactcatcacgatggtt	C	cast	269
	atcactcatagcccaga	D	129/	270
	atcacttatagcccaga	F	spre	271
	atcactcatatcccaga	E	cast	272
107	catcttaccagcattga	A	129/cast/spre	273
	catcttactagcattga	B	b6/fvbn/f1	274
108	agtcagccggctctggc	A	129/b6/f1	275
	agtcagccagctctggc	B	fvbn/f1	276
	gggtaggagtggggatgag	C	129/	277
	gggcaggagtgggggtgag	E	spre	278
	gggtaggagtgggggtgag	D	cast	279
109	tcagtattgttcttctc	A	129/f1/spre	280
	tcagtatttttcttctc	B	b6/fvbn/f1/cast	281
110	agcagagactgagctcg	A	129/	282
	agcagagaccgagctcg	B	b6/fvbn/f1	283

	acaggggtcgattcgtc	C	129/b6/fvbn/f1/cast	284
	acagggatcgattcgtc	E	spre	285
	acaggggtcgtttcgtc	D	f1	286
111	tcccaaagcattcaagg	A	129/b6/f1	287
	tcccaaagtattcaagg	B	fvbn/f1	288
	gaccagggttaatgact	C	129/b6	289
	gaccagggctaatgact	D	cast/spre	290
112	ctattaacagagtcgag	A	129/b6/f1	300
	ctattaacggagtcgag	B	fvbn	301
	gtgatactggatgtctg	C	129/b6	302
	gtgataccg-atgtctgg	D	cast/spre	303
113	ctctctcgatagtctaa	A	129/f1	304
	ctctctcgctagtctaa	B	b6/fvbn/f1/cast	305
	tctctcgatagtctaat	C	129/	306
	tctctcgctggtctaat	D	cast	307
114	agatgcaaaattcttag	A	129/	308
	agatgcacagttcttag	B	b6/fvbn/f1	309
115	ggaaaatgctcaggtag	A	129/f1/cast/spre	310
	ggaaaatgttcaggtag	B	b6/fvbn	311
116	tctgggcagagtgcagg	A	129/	312
	tctgggcagcgtgcagg	B	b6/fvb/f1	313
117	tatggaacggttgcttc	A	129/fvb	314
	tatggaactgttgcttc	B	b6/f1	315
	aagcctggtacccgctg	C	129/cast	316
	aagcctggcacccgctg	D	spre	317
118	cattcttctttttctga	A	129/	318
	cattcttcgttttctga	B	b6/fvbn/f1/cast/spre	319
	ctgcaggcttgtctgtg	C	129/CAST	320
	ctgcaggtttgtctgtg	D	spre	321
119	tgccatttcctataaca	A	129/f1	322
	tgccatttgctataaca	B	b6/fvbn	323
120	ccgccacacccgctcct	A	129/b6	324
	ccgccacagccgctcct	B	fvbn/f1	325
121	caaataatgctagttat	A	129/b6/f1	326
	caaataatgttagttat	B	fvbn	327

122	ggatgttgacacgctac	A	129/fvbn/f1	328
	ggatgttgtcacgctac	B	b6/f1	329
	catgtgtc-caacgccat	C	129/	330
	catgtgtcacaacgcca	D	cast/spre	331
123	aaaggggccttaaagga	A	129/fvbn/f1	332
	aaaggggctttaaagga	B	b6	333
	tgaaaagttcttttcat	C	129/cast	334
	tgaaaagtacttttcat	D	spre	335
124	cctctctatgtgtgagc	A	129/b6/f1	336
	cctctctacgtgtgagc	B	fvbn	337
	gaagttttaggagattct-t	C	129/	338
	gaagatttaggagagtctc	D	spre	339
125	agggatgtattttgtta	A	129/fvbn/f1	340
	agggatgtgttttgtta	B	b6	341
	acaattcaaatgtatat	C	129/cast	342
	acaattcatatgtatat	D	spre	343
126	cttgcctaacctgcaca	A	129/b6/f1	344
	cttgcctagcctgcaca	B	fvbn	345
	caacagc---acctcatatc	C	129/b6/cast	346
	acagcggtgcctcgtat	D	spre	347
127	actcacagtgtcagggc	A	129/fvbn/f1/spre	348
	actcacagcgtcagggc	B	b6/cast	349
128	ggctgctcctgtgtgtctg	A	129/fvbn/f1/cast	350
	ggctcttcctgtgtgtctg	B	b6	351
	ggctgctcctgtgtttctg	C	spre	352
129	aatagatgcccttctga	A	129/f1	353
	aatagatgccctcttga	B	b6/fvbn	354
	aatcgatgcccttctga	C	spre	355
130	ttggtctagcaggtagc	A	129/fvbn/f1	356
	ttggtctaccaggtagc	B	b6	357
	agccttggctcttaaaa	C	129/cast	358
	agccttggttcttaaaa	D	spre	359
131	agtctctggcgcctttg	A	129/fvbn/f1/cast/spre	360

	agtctctgccgcctttg	B	b6	361
132	tagcaggaggcacagctta	A	129/	362
	aagcaggaggcacaactta	B	b6	363
	aagcaggaggcacagctta	C	fvb/f1/CAST	364
	tagcaggaggcacagcttg	D	spre	365
133	aggagagaccggactcc	A	129/fvb/f1	366
	aggagagagcggactcc	B	b6	367
134	tacaagtcatccttcct	A	129/b6/f1	368
	tacaagtcgtccttcct	B	fvbn/f1	369
	atacctccctcagacaa	C	129/cast	370
	atacctcc-tcagacaag	D	spre	371
135	aaacaaacaaacaaacc	A	129/b6/f1/cast/spre	372
	aaacaaaccaacaaacc	B	fvbn	373
	gtgcgccaccatgacca	C	129/cast	374
	gtgcgccatcatgacca	D	spre	375
136	ggctttcccattagtgg	A	129/	376
	ggctttcctattagtgg	B	b6/fvbn/f1	377
	ccctcacctctctctca	C	129/cast	378
	ccctcacccctctctca	D	spre	379
137	aatctctcgcgttcatt	A	129/fvbn/f1	380
	aatctctcacgttcatt	B	b6	381
138	aatgataccgatcctta	A	129/f1	382
	aatgatacagatcctta	B	b6/fvbn	383
	ataaaactgcattcgtg	C	129/b6	384
	ataaaactacattcgtg	D	cast/spre	385
B1Musch	AGTTCCAGGACAGCCAGG			
201	atatctccgactttgaa	A	129/cast	386
	atatctccaactttgaa	B	b6/fvb/f1/spre	387
	tggccctgcagagtctg	C	129-Cast	388
	tggctctgcagag-ctgg	D	Spre	389
202	caatggatc---aaagatgc	A	129-FVB-F1	390
	atggatcaacaaagatg	B	B6	391
	gctgcctc--aaggtataa	C	129/b6	392
	ctgcctcttaaggtata	D	cast/spre	393

	agtttggtcccctggac	C	129/FVB/B6-F1-Cast	430
	agtttggtttcctggac	D	Spre	431
214	tatagcttcatgtaaaa	A	129/fvb/f1/cast/spre	432
	tatagctttatgtaaaa	B	b6	433
215	tttttttt-attattgaa	A	129	434
	ttttttttttattattga	B	B6-FVB-F1	435
	actcattgccaatttaa	C	129	436
	actcattcagaatttaa	D	spre/CAST	437
216	atgcgtaatgggggcta	A	129	438
	atgcgtaacgggggcta	B	b6/fvb/f1/cast/SPRE	439
	ataattgctcttttaaa	C	129/b6/fvb/f1/cast	440
	gtaattgctcttttaaa	D	spre	441
217	tctgattagtgatggat	A	129-F1	442
	tctgatta-tgatggatt	B	B6	443
	agcagagtgtctcgtaa	C	129	444
	agcagagtatctcgtaa	D	spre/CAST	445
218	gctggcagatatcggta	A	129/b6/f1	446
	gctggcaggtatcggta	B	fvb/cast	447
219	aactgcaatgaccagca	A	129-B6	448
	aactgcaacgaccagca	B	FVB-F1	449
	gctggtcattgcagttt	C	129	450
	gttggtcgttacagttt	D	spre	451
	gctggtcgttgcagttt	E	cast	452
220	gctggcagatatcggta	A	129-B6-F1	453
	gctggcaggtatcggta	B	FVB	454
	atagaaagtccaccgtc	C	129/cast	455
	atagaaagcccaccgtc	D	spre	456
221	ttagtgaccgtgtaaac	A	129/b6/f1	457
	ttagtgactgtgtaaac	B	fvb	458
	ggggaggagctttgttc	C	129-Cast	459
	ggggaggatctttgttc	D	Spre	460
222	ggcctggacacaaaagc	A	129/fvb/f1	461

203	acctatggctcctcatc	A	129/b6/f1	394
	acctatggttcctcatc	B	fvb	395
	tcttctcccctgcttta	C	129-Cast	396
	tcttctcac-tgctttag	D	Spre	397
204	ccgc-ataaaaagctgag	A	FVB-F1	398
	ccgccataaaa-gctgag	B	B6-F1	399
	agaatatagggtttttt	C	129/cast	400
	tagaatacag--ttttttt	D	spre	401
205	agagttgctgtgcaggg	A	129/b6/f1	402
	agagttgccgtgcaggg	B	fvb/cast	403
	agagttgcagtgcaggg	C	spre	404
206	taagcagtgttcttggc	A	129-B6-F1	405
	taagcagtattcttggc	B	FVBN	406
	tcttctcccctgcttta	C	129/cast	407
	tcttctcac-tgctttag	D	spre	408
207	ttttttttattattga	A	129/fvb/f1	409
	tttttttt-attattgaa	B	b6	410
	tgtggtacgcacatctg	C	129-Cast	411
	tgtggtacacacatctg	D	Spre	412
208	agactcttagacttctg	A	129/f1	413
	agactcttaggcttctg	B	b6/fvb/f1	414
	agactcataagcttctg	C	spre	415
	agactcttaggcttctg	D	cast	416
209	cacgtacccgaacgtga	A	129-B6	417
	cacgtacctgaacgtga	B	FVB-F1	418
	attacggtttgtcgtca	C	129/CAST	419
	attacggttggtcgtca	D	spre	420
210	ccaagatacgaaaccag	A	129/f1/cast/spre	421
	ccaagatatgaaaccag	B	b6	422
211	tgcaatgaccagcaacc	A	129/b6	423
	tgcaacgaccagcaacc	B	fvb/f1/cast	424
	tgtaacgaccaacaact	C	spre	425
212	tctaaagggaaagatgg	A	129-FVB	426
	tctaaagg-aaagatgga	B	B6-F1	427
213	ctggactcatacataca	A	129-FVB-F1	428
	ctggactcgtacataca	B	B6-F1-Cast/SPRE	429

	ggcctggaaacaaaagc	B	b6	462
	ccctttctagtattgt	C	129	463
	ccctttccagtattgt	D	Cast-Spre	464
223	gaattggttttaggaat	A	129-F1-Cast-Spre	465
	gaattggtattaggaat	B	B6	466
224	acccagctttccatggt	A	129/f1	467
	acccagctctccatggt	B	b6/fvb/CAST	468
225	tcacgttcgggtacgtg	A	129/b6/f1	469
	tcacgttcaggtacgtg	B	fvb/f1	470
	tgccttccggttggcaa	C	129-Cast	471
	tgccttccagttggcaa	D	Spre	472
226	ttttatcatacaattgc	A	129-F1	473
	ttttatcagacaattgc	B	B6-FVB-F1	474
227	atcttctcttctttgag	A	129/f1	475
	atcttctcctctttgag	B	b6/fvb	476
	cagtcctctgctttctc	C	129-Cast	477
	cagtcctcagctttctc	D	Spre	478
228	ccaagatacgaaaccag	A	129/f1/spre	479
	ccaagatatgaaaccag	B	b6	480
229	ggtattcaagggttact	A	129/cast/spre	481
	ggtattca-gggttactg	B	b6/fvb 1bp del.	482
230	acctatggctcctcatc	A	129/b6/f1/cast	483
	acctatggttcctcatc	B	fvb	484
231	ttttatcatacaattgc	A	129/f1	485
	ttttatcagacaattgc	B	b6/fvb	486
232	aaccagggcttaagtct	A	129	487
	aaccagggattaagtct	B	b6/fvb/f1	488
	cagaaaaacagatatac	C	129-B6-FVB-F1	489
	cagaaaaagagatatac	D	Spre	490
234	tctgagcgtgagtgctg	A	129/fvb	491
	tctgagcgcgagtgctg	B	b6/f1/cast/spre	492
	acctcagaagcggaggt	C	129-B6-FVB-F1	493
	acctcggaaggggaggt	D	Spre	494
	acctcggaagcggaggt	E	Cast	495

235	taactcgatcgctatca	A	129-B6-F1	496
	taactcgcttgctatca	B	FVBN-Cast	497
	taactcgctcgctatca	C	Spre	498
236	gaatttctcaacttctt	A	129/fvb/f1/spre	499
	gaatttctgaacttctt	B	b6/f1	500
237	caggggtccccaatttg	A	129/f1/SPRE	500
	caggggtctccaatttg	B	b6/fvb	501
238	ttttgctgtgc-aggcta	A	129-B6-F1	502
	ttttactgtgccaggct	B	FVB	503
	gacagccctgtctcaaa	C	129/cast	504
	agagaaaccctgtctca	D	spre	505
239	gcaccggtctgagcagt	A	129/f1	506
	gcaccggtttgagcagt	B	b6/fvb/f1	507
	ccgtgcccctgaacaat	C	129-B6-FVB-F1-Cast	508
	ccgtgcccttgaacaat	D	Spre	509
240	tcacgttcgggtacgtg	A	129/b6/f1	510
	tcacgttcaggtacgtg	B	fvb/f1	511
	tgattcgctgggactct	C	129-Cast	512
	tgattcgccgggactct	D	Spre	513
241	ttgatatccgaggcctt	A	129/b6/fvb/f1	514
	ttgatatctgaggcctt	B	f1/CAST/SPRE	515
242	tccctgggccaagcata	A	129/b6/fvb	516
	tccctgggtcaagcata	B	f1	517
243	ttatggctgaggatcac	A	129-B6-F1-Cast	518
	ttatggctgcggatcat	B	FVB	519
	ttatggcaggggatcac	C	Spre	520
244	ctctctgcgctgaagca	A	129/b6	521
	ctctctgctctgaagca	B	fvb/f1	522
	agatacagagatgtgtt	C	129-B6-FVB-F1	523
	agatactgaggtgtgtt	D	Spre	524
245	cgacatctggcagatgt	A	129/f1	525
	cgacatctagcagatgt	B	b6/fvb	226
	gtcacaaatagtatttc	C	129/cast	527
	gtcacaaagagtatttc	D	spre	528
246	aaggtgtgtgcgtgtgt	A	129/f1	529

	aaggtgtgcgcgtgtgt	B	fvb	530
247	agtctttttttcctga	A	129-B6-FVB	531
	tagtc-tttttttt-cctgaa	B	F1	532
248	caggctgtgggaggctt	A	129/b6/f1	533
	caggctgcggaaggctt	B	fvb	534
	ctgtaagtcattcaata	C	129-B6-FVB-F1-Cast	535
	ctgtaagtaattcaata	D	Spre	536
249	caggggtccccaatttg	A	129/f1	5367
	caggggtctccaatttg	B	b6/fvb	538
250	gactcatggccgccttg	A	129	539
	gactcattgccgcctgg	B	B6-FVB-F1	540
	gactcctggccgcctgg	C	F1	541
	gactcctggctgcctgg	D	Spre	542
	gactcctggccgcctgg	E	Cast	543
251	acagggga-ggaaggaag	A	129	544
	acaggggaaggaaggaa	B	b6/fvb/f1	545
252	ttgatatagattgattc	A	129/b6/f1	546
	ttgatatatattgattc	B	fvb/f1	547
	atagaacagcaaagtaa	C	129-B6-FVB-F1-Cast	548
	atagaacaacaaagtaa	D	Spre	549
253	aacaagcatctatggat	A	129/fvb/f1	556
	aacaagcacctatggat	B	b6	551
DOP				
300	gagcaggttaagcgatg	A	129/	552
	gagcaggtgaagcgatg	B	B6	553
301	ggcttccagcttgattc	A	129/	554
	ggcttccaacttgattc	B	B6	555
302	agatagggatgaatccc	A	129/	556
	agataggggtgaatccc	B	B6	557
303	tcattcaccgtttattg	A	129/	558
	tcattcactgtttattg	B	B6	559
304	ctgacatactgcttagg	A	129/	560
	ctgacatattgcttagg	B	B6	561
305	ctaggaaagcctaaatt	A	129/	562

	ctaggaaaacctaaatt	B	B6	563
306	atgtcaggattttaaga	A	129/	564
	atgtcagggttttaaga	B	B6	565
307	ggtttccaattggaaag	A	129/	566
	ggtttccagttggaaag	B	B6	567
308	cgaggagtgcaaagcga	A	129/	568
	cgaggagtccaaagcga	B	B6	569
309	tgtgtgtgtctgtct	A	129/	570
	tgtgtgtgcgtctgtct	B	B6	571
310	gcaagatgcagctgcat	A	129/	572
	gcaagatgtagctgcat	B	B6	573
311	gctggggctattctgta	A	129/	574
	gctggggccattctgta	B	B6	575
312	caataacggacctgcct	A	129/	576
	caataacgaacctgcct	B	B6	577
313	tagcctctctacatagg	A	129/	578
	tagcctctgtacatagg	B	B6	579

Other SNPs identified using the BJ1 DOP-PCR Primer include:

SNPs present within DOP-PCR using primer BJ1						
	Genotype of CEPH individuals:					
ASO name	ASO sequence	12-01	104-01	884-01	1331-01	SEQ ID#
3A-G	CATCTATAGGTTCACTT	GT	TT	TT	TT	580
3A-T	CATCTATATGTTCACTT					581
5A-C	GCCAACAACATTGAGAG	GG	CG	GG	GG	582
5A-G	GCCAACAAGATTGAGAG					583
7A-C	GGGTCGTGCGTCCCCCT	TT	CT	TT	TT	584
7A-T	GGGTCGTGTGTCCCCCT					585
9A-A	ATTGTCTCACATTTCTT	AA	GG	AA	AA	586
9A-G	ATTGTCTCGCATTTCTT					587
12A-C	GGTGTGGTCGCAGAAGG	CC	CC	CT	CT	588
12A-T	GGTGTGGTTGCAGAAGG					589
15A-A	TCATTGCCACACTTGAA	AA	GG	AA	GG	590
15A-G	TCATTGCCGCACTTGAA					591
20A-A	ATCTGTCTACAATGATC	AG	GG	AA	AG	592
20A-G	ATCTGTCTGCAATGATC					593
22A-A	GGCTGGGCACAGTGGCT	AA	GG	AA	AA	594
22A-G	GGCTGGGCGCAGTGGCT					595
34A-A	CAGCCTGGAGAACAAGT	CC	CC	CC	AC	596
34A-C	CAGCCTGGCGAACAAGT					597
39A-C	TTTGACACCCGGAAGCT	CT	CC	CC	CC	598
39A-T	TTTGACACTCGGAAGCT					599
40A-C	CTGCCTTTCATACTGCC	CT	TT	CT	TT	600
40A-T	CTGCCTTTTATACTGCC					601
40B-C	ACAATAGACGTTCCCCG	TT	CT	TT	CT	602
40B-T	ACAATAGATGTTCCCCG					603
41A-A	GGTGTTTGATTTGTACT	CC	AC	CC	CC	604
41A-C	GGTGTTTGCTTTGTACT					605
42A-A	TCCAACTCAAAAAATGT	AT	AA	AT	AT	606
42A-T	TCCAACTCTAAAAATGT					607
44A-C	GGGCCGCTCACAGTCCA	CC	CT	CC	CC	608
44A-T	GGGCCGCTTACAGTCCA					609
44B-C	GCATGGCTCGTGGGTTT	CT	CT	TT	CT	610
44B-T	GCATGGCTTGTGGGTTT					611
46A-G	GTTGGGAAGTGGAGCGG	GG	TT	GG	TT	612
46A-T	GTTGGGAATTGGAGCGG					613
50A-A	AAGGGATGAGGATGTGA	AG	AA	AA	AG	614
50A-G	AAGGGATGGGGATGTGA					615
50B-A	TCCTCGAGAGCTTTGCT	AG	AG	AA	AG	616
50B-G	TCCTCGAGGGCTTTGCT					617
51A-C	TGACAATGCGTGCCCAA	CT	CC	CC	CC	618
51A-T	TGACAATGTGTGCCCAA					619
53A-A	TCCATGTCATAGATTTC	AG	AA	AA	AA	620
53A-G	TCCATGTCGTAGATTTC					621

66A-A	TGGAGGACAGTGGAGGG	TT	TT	TT	AT	622
66A-T	TGGAGGACTGTGGAGGG					623
69A-C	ACCCATTTCCTGAAAAT	TT	CT	TT	TT	624
69A-T	ACCCATTTTCTGAAAAT					625
71A-G	CTGAGTTCGGCACTGCT	TT	GG	GG	TT	626
71A-T	CTGAGTTCTGCACTGCT					627
71B-G	ACCAGTTTGGCTCAAAG	GG	TT	TT	GG	628
71B-T	ACCAGTTTTGCTCAAAG					629
72A-A	CCAATCAGAACGTGCAG	AA	GG	GG	AA	630
72A-G	CCAATCAGAGCGTGCAG					631
73A-A	ACCCACACAGACACTGC	AA	AT	TT	AT	632
73A-T	ACCCACACTGACACTGC					633
81A-C	GGACAAAGCGCTGGTGT	TT	CT	CC	CT	634
81A-T	GGACAAAGTGCTGGTGT					635
81C-C	AGCTGGTCCCCCTMCCC	TT	CT	CC	CC	636
81C-T	AGCTGGTCTCCCTMCCC					637
90A-A	GGTGTAGTAAGCACAGC	AA	AA	AC	AA	638
90A-C	GGTGTAGTCAGCACAGC					639
91A-C	AGCGAACACGGGGGAAA	CC	CC	TT	CC	640
91A-T	AGCGAACATGGGGGAAA					641
98D-A	GTGACAGCACCAAACTT	GG	AG	GG	GG	642
98D-G	GTGACAGCGCCAAACTT					643
101A-C	GTCTGTTGCTGTTATTT	TT	TT	TT	CT	644
101A-T	GTCTGTTGTTGTTATTT					645
111A-A	ACCAGCATAGCCCAGAG	GG	GG	GG	AG	646
111A-G	ACCAGCATGGCCCAGAG					647
111B-A	CGTAGGAGACAAGACCT	GG	GG	GG	AG	648
111B-G	CGTAGGAGGCAAGACCT					649
117A-A	CTCTGCTGAATCTCCCA		GG	GG	AG	650
117A-G	CTCTGCTGGATCTCCCA					651
124A-A	AAGCAAAGACTGATTCA	TT	AT	TT	TT	652
124A-T	AAGCAAAGTCTGATTCA					653
125A-A	AGGCAGCTAGAGGGAGA	CC	AA	AC	AA	654
125A-C	AGGCAGCTCGAGGGAGA					655
130C-C	TTCCATTCCGTTCAATT	TT	TT	TT	CC	656
130C-T	TTCCATTCTGTTCAATT					657
130D-C	TATTGTTACTGATTTTG	CT	CT	CT	TT	658
130D-T	TATTGTTATTGATTTTG					659
136A-A	GAGCTTTCAGAGGCTGA	AA	AG	AG	AG	660
136A-G	GAGCTTTCGGAGGCTGA					661
137A-A	GGGGGAAGATATGGAGT	GG	AG	AA	AG	662
137A-G	GGGGGAAGGTATGGAGT					663
143A-C	CATGGCCTCGTGGGTTT	TC	TC	TT	TC	664
143A-T	CATGGCCTTGTGGGTTT					665
147B-A	GGGKAGGGAGACCAGCT	AA	AG	GG	GG	666
147B-G	GGGKAGGGGGACCAGCT					667

147C-A	GCAGTGTCAGTGTGGGT	TT	AT	AA	AT	668
147C-T	GCAGTGTCTGTGTGGGT					669
147D-A	ACACCAGCACTTTGATC	AA	AG	GG	AG	670
147D-G	ACACCAGCGCTTTGATC					671
151A-A	CCTTCTGCAACCACACC	GG	GG	AG	AG	672
151A-G	CCTTCTGCGACCACACC					673
163A-A	AAATTCGCAGGAGCCGA	GG	AG	GG	GG	674
163A-G	AAATTCGCGGGAGCCGA					675
164B-A	AGGTCTAGACGCTCACC	AG	GG	AG	GG	676
164B-G	AGGTCTAGGCGCTCACC					677
164C-A	GGAGGAACACTTCAAAC	GG	AG	GG	GG	678
164C-G	GGAGGAACGCTTCAAAC					679
170A-A	TTTGTGCTATACCTTGA	AA	AG	AG	AG	680
170A-G	TTTGTGCTGTACCTTGA					681
179A-C	ATGATGCACACACCCTC	CT	CC	TT	CC	682
179A-T	ATGATGCATACACCCTG					683
181B-C	TATTGCTCCGCCTCCTC	CT	TT	CC	TT	684
181B-T	TATTGCTCTGCCTCCTC					685
181D-C	CTCAGAGACTGTGTGCC	CG	CC	CC	CC	686
181D-G	CTCAGAGAGTGTGTGCC					687
187A-C	ATCTTCTGCGTCACTCA	CT	CT	CC	CC	688
187A-T	ATCTTCTGTGTCACTCA					689
187B-A	CAGCATCTAGTAACCAC	AG	AA	GG	AG	690
187B-G	CAGCATCTGGTAACCAC					691
190A-C	ATTAGTGCCAAATACAT	CC	CC	CT	CT	692
190A-T	ATTAGTGCTAAATACAT					693
195B-A	TGCTCCACAGCAGCCGT	AT	TT	TT	TT	694
195B-T	TGCTCCACTGCAGCCGT					695
196A-A	TAGGGGAGAATCTGTTT	CC	AC	AC	AA	696
196A-C	TAGGGGAGCATCTGTTT					697

The invention also encompasses a composition comprising a plurality of RCGs immobilized on a surface, wherein the RCGs are composed of a plurality of DNA fragments, each DNA fragment including a $(N)_x$-TARGET polynucleotide structure as described above, i.e., wherein the TARGET portion is identical in all of the DNA fragments of each RCG, the portion includes at least 7 nucleotide residues, wherein x is an integer from 0 to 9, and wherein each N is any nucleotide residue. Preferably the TARGET portion includes at least 8 nucleotides residues.

In other aspects, the invention includes a method for performing DOP-PCR. The prior art DOP-PCR technique was originally developed to amplify the entire genome in cases where DNA was in short supply. This method is accomplished using a primer set wherein each primer has an arbitrarily selected six nucleotide residue portion, at its 3' end. The complexity of the resultant product is extremely high due to the short length and results in amplification of the genome. By increasing the length of the arbitrarily selected of the DOP-PCR primer from 6 nucleotides to 7, and preferably 8, or more nucleotide residues the complexity of the genome is significantly reduced.

Examples

Example 1: Identification and isolation of SNPs

High allele frequency SNPs are estimated to occur in the human genome once every kilobase or less (Cooper et al., 1985). A method for identifying these SNPs is illustrated in

Figure 1. As shown in Figure 1, inter-Alu PCR was performed on genomes isolated from three unrelated individuals. The PCR products were cloned, and a mini library was made for each of the 3 individuals. The library clone inserts were PCR-amplified and spotted on nylon filters. Clones were matched by hybridization into two sets of identical clones from each individual, for a total of 6 clones per matched clone set. These sets of clones were sequenced, and the sequences were compared in order to identify SNPs. This method of identifying SNPs has several advantages over the prior art PCR amplification methods. For instance, a higher quality sequence is obtained from cloned DNA than is obtained from cycle sequencing of PCR products. Additionally, every sequence represents a specific allele, rather than potentially representing a heterozygote. Finally, sequencing ambiguities, Taq polymerase errors, and other source of sequence error particular to one representation of the sequence are reduced by application of an algorithm which requires that the same variant sequence be present in at least 2 of the 6 clones sampled.

In general, the Alu PCR method for identifying SNPs can be performed using genomic DNA obtained from independent individuals, unrelated or related. Briefly, Alu PCR is performed which yields a product having an estimated complexity of approximately 100 different single copy genomic DNA sequences and an average sequence length of between about 500 base pairs and 1 kilobase pairs. The PCR products are cloned, and a mini library is made for each individual. Approximately 800 clones are selected from each library and transferred into a 96-well dish. Filter replicas of each plate are hybridized with PCR probes from individual clones selected from one of the libraries in order to create a matched clone set of 6 clones, 2 from each individual. Many sets of clones can be isolated from these libraries. The clones can be sequenced and compared to identify SNPs.

Methods

An Alu primer designated primer 8C was designed to produce an Alu PCR product having a complexity of approximately 100 independent products. Primer 8C (having the nucleotide sequence CTT GCA GTG AGC CGA GATC; SEQ ID NO: 3) is complementary with base pairs 218-237 of the Alu consensus sequence (Britten et al., 1994). In order to reduce the complexity of the product, however, the last base pair of the primer was selected to correspond to

base pair 237 of the consensus sequence, a nucleotide which has been shown to be highly variable among Alu sequences. Primer 8C therefore produces a product having complexity lower than that produced using Alu primers which match a segment of the Alu sequence in which there is little variation in nucleotide sequence among Alu family members.

Preliminary experiments were conducted to estimate the complexity of the product produced by Alu PCR reaction with primer 8C on the CEPH Mega Yacs. These preliminary experiments confirmed that primer 8C produced a lower number of Alu PCR products than other Alu PCR primers closely matching less variable sequences in the Alu consensus.

Three libraries of Alu PCR products were produced from inter-Alu PCR reactions involving genomic DNA derived from three unrelated CEPH individuals designated 201, 1701, and 2301. The reactions were performed at an annealing temperature of 58°C for 32 cycles using the 8C Alu primer. Each set of PCR reaction products was purified by phenol:chloroform extraction followed by ethanol precipitation. The products were shotgun cloned into the T-vector pCR2.1 (Invitrogen); electroporated into *E. coli* strain DH10B Electromax ampicillin-containing LB agar plates. 768 colonies were picked from each of the three libraries into eight 96-well format plates containing LB+ ampicillin and grown overnight. The following day, an equal volume of glycerol was added and the plates were stored at -80°C. An initial survey of the picked clones indicated an average insert size of between 500 base pairs and 1 kilobase pair.

To identify matching clones in each library, 1 microliter of an overnight culture made from each library plate well was subjected to PCR amplification using vector-derived primers. Amplified inserts were spotted onto Hybond ™ N+ filters (Amersham) using a 96-pin replicating device such that each filter had 384 products present in duplicate. The DNA was subjected to alkali denaturation by standard methods and fixed by baking at 80°C for 2 hours. Individual inserts derived from the library were radiolabeled by random hexamer priming and used as probes against the three libraries (6 filters per probe). Hybridization was carried out overnight at 42°C in buffer containing 50% formamide as described in Sambrook et al. The following day, the filters were washed in 2X standard saline citrate (SSC), 0.1% SDS at room temperature for 15 minutes, followed by 2 washes in 0.1X SSC, 0.1% SDS at 65°C for 45 minutes each. The filters were then exposed to Kodak X-OMAT X-ray film overnight.

Results

Figure 2 shows the data obtained for identification of SNPs. The results of the gel electrophoresis of inter-Alu PCR genomic DNA products prepared using the 8C primer is shown in Figure 2A. Mini libraries were prepared from the Alu PCR genomic DNA products. Colonies were picked from the libraries, and inserts were amplified. The inserts were separated by gel electrophoresis to demonstrate that each was a single insert. The gel is shown in Figure 2B. Once the individual amplified inserts were spotted on Hybond™ N+ filters, the inserts were radiolabeled by random hexamer primary and used as probes of the entire contents against the three mini libraries. One of the filters, having 2 positive or matched clones, is shown in Figure 2C.

The results of screening 330 base pairs of genomic DNA by the matched clone method led to the identification of 6 SNPs, 4 in single copy DNA, 2 in the flanking Alu sequence. These observations were consistent with the projected rate of SNP currents of 1 high frequency SNP per 1,000 base pairs or less. The single copy SNPs identified are presented below in Table I.

Table I

CEPH Individual	1	2	3	4
201	taagtGtacaa (SEQ ID NO. 5)	cccacGgagaa (SEQ ID NO. 7)	aattgCttccc (SEQ ID NO. 9)	aaattCaatgt (SEQ ID NO. 11)
	taagtGtacaa (SEQ ID NO. 5)	cccacGgagaa (SEQ ID NO. 7)	aattgCttccc (SEQ ID NO. 9)	aaattCaatgt.. (SEQ ID NO. 11)
1701	taagtAtacaa (SEQ ID NO. 6)	cccacAgagaa (SEQ ID NO. 8)	aattgCttccc (SEQ ID NO. 9)	aaattCaatgt.. (SEQ ID NO. 11)
	taagtGtacaa (SEQ ID NO. 5)	cccacGgagaa (SEQ ID NO. 7)	aattgTttccc (SEQ ID NO. 10)	aaattCaatgt.. (SEQ ID NO. 11)
2301	taagtGtacaa (SEQ ID NO. 5)	cccacAgagaa (SEQ ID NO. 8)	aattgCttccc (SEQ ID NO. 9)	aaattAaatgt.. (SEQ ID. NO. 12)

	taagtGtacaa (SEQ ID NO. 5)	cccacGgagaa (SEQ ID NO. 7)	aattgTttccc (SEQ ID NO. 10)	aaattCaatgt.. (SEQ ID NO. 11)

To verify the identities of the SNPs shown in Table I, specific primers were synthesized which permitted amplification of each single copy locus. Cycle sequencing was then performed on PCR products from each of the three unrelated individuals, and the site of the putative SNP was examined. In all cases, the genotype of the individual derived by cycle sequencing was consistent with the genotype observed in the matched clone set.

Example 2: Allele-specific oligonucleotide hybridization to Alu PCR SNPs

Methods

Inter-Alu PCR was performed using genomic DNA obtained from 136 members of 8 CEPH families (numbers 102, 884, 1331, 1332, 1347, 1362, 1413, and 1416) using the 8C Alu primer, as described above. The products from these reactions were denatured by alkali treatment (10-fold addition of 0.5 M NaOH, 2.0 M NaCl, 25 mM EDTA) and dot blotted onto multiple Hybond ™ N+ filters (Amersham) using a 96-well dot blot apparatus (Schleicher and Schull). For each SNP, a set of two allele-specific oligonucleotides consisting of two 17-residue oligonucleotides centered on the polymorphic nucleotide residue were synthesized. Each filter was hybridized with 1 picomole ^{32}P-kinase labeled allele-specific oligonucleotides and a 50-fold excess of non-labeled competitor oligonucleotide complementary to the opposite allele (Shuber et al., 1993). Hybridizations were carried out overnight at 52°C in 10 mL TMAC buffer 3.0 M TMAC, 0.6% SDS, 1 mM EDTA, 10 mM $NaPO_4$, pH 6.8, 5X Denhardt's solution, 40 micrograms/milliliter yeast RNA). Blots were washed for 20 minutes at room temperature in TMAC wash buffer (3 M TMAC, 0.6% SDS, 1 mM EDTA, 10 mM Na_3PO_4 pH 6.8) followed by 20 minutes at 52°C (52°C-52°C is optimal). The blots were then exposed to Kodak X OMAT AR X-ray film for 8-24 hours and genotypes were determined by the hybridization pattern.

Results

The results of the genotyping and mapping are shown in Figure 3. In order to determine

the map location of the SNP, the genotype data determined from CEPH families number 884 and 1347 were compared to the CEPH genotype database version 8.1 (HTTP:\\www.cephb.fr/cephdb/) by calculating a 2 point lod score using the computer software program MultiMap version 2.0 running on a Sparc Ultra I computer. This analysis revealed a linkage to marker D3S1292 with a lod score of 5.419 at a theta value of 0.0. To confirm this location, PCR amplification of the CCRSNP1 marker was performed on the Gene Bridge 4 radiation hybrid panel (Research Genetics). This analysis placed marker CCRSNP1 at 4.40 cR from D3S3445 with a lod score greater than 15.0. Integrated maps from the genetic location database (Collins et al., 1996) indicated that the locations of the markers identified by these two independent methods are overlapping. These results support the mapping of even low frequency polymorphisms by two point linkage to markers previously established on CEPH families.

Of the dot blots performed on each CEPH family PCR, two families were informative at this SNP locus, namely families number, 884 and 1347. The dot blot is shown in Figure 3A. Lines are drawn around signals representing CEPH family 884 on the dot blot shown in Figures 3A and 3B. Allele-specific oligonucleotide hybridizations were performed on the filters shown in Figures 3A and 3B under TMAC buffer conditions with G allele-specific oligonucleotide (Figure 3A) and A allele-specific oligonucleotide (Figure 3B). The pedigree of CEPH family number 884 with genotypes as scored from the filter shown in Figures 3A and 3B is shown in Figure 3C. The DNA was not available for one individual in this pedigree, and that square is left blank. Mapping of CCRSNP1 was performed by two independent methods. First, genotype data from informative CEPH families numbers 884 and 1347 were compared to the CEPH genotype database version 8.1 by calculation of a 2 point lod score. Secondly, PCR amplification of the CCRSNP1 marker was performed on the Gene Bridge 4 radiation hybrid panel. The highest lod scores determined by these analyses were D3S1292 and D3S3445, respectively, as shown in Figure 3D.

The percentage of SNPs detected using the above-described methods is dependent on the number of chromosomes sampled, as well as the allele frequency.

Example 3: Confirmation of SNP identity

Allele-specific oligonucleotides are synthesized based on standard protocols (Shuber et al., 1997). Briefly, polynucleotides of 17 bases centering on the polymorphic site are synthesized for each allele of a SNP. Hybridization with DNA dots of IRS or DOP-PCR products affixed to a membrane were performed, followed by hybridization to end labeled allele-specific oligonucleotides under TMAC buffer conditions. These conditions are known to equalize the contribution of AT and GC base pairs to melting temperature, thereby providing a uniform temperature for hybridization of allele-specific oligonucleotides independent of nucleotide composition.

Using this methodology, genotypes of CEPH progenitors and their offspring are determined. The Mendelian segregation of each SNP marker confirms its identity as a SNP marker and accrued estimate of its relative allele frequency, hence, its likely usefulness as a genetic marker. Markers which yield complex segregation patterns or show very low allele frequencies on CEPH progenitors are set aside for future analysis, and remaining markers are further characterized.

Example 4: Development of detailed information on map position and allele frequency for each SNP

Two complementary methods are used to establish genetic map position for each marker. Each marker is genotyped on a number of CEPH families. The result is compared, using MultiMap (Matise et al., 1993, as described above) or other appropriate software, against the CEPH database to determine by linkage the most likely position of the SNP marker.

Allele frequencies are determined by hybridization with the standard worldwide panel which U.S. NIH currently is making available to researchers for standardization of allele frequency comparison. Allele-specific oligonucleotide methodology used for genetic mapping is used to determine allele frequency.

Example 5: Development of a system for scoring genotype using SNPs

After the identification of a set of SNPs, automated genotyping is performed. Genomic DNA of a well-characterized set of subjects, such as the CEPH families, is PCR- amplified using

appropriate primers. These DNA samples serve as the substrate for system development. The DNA is spotted onto multiple glass slides for genotyping. This process can be carried out using a microarray spotting apparatus which can spot greater than 1,000 samples within a square centimeter area or more than 10,000 samples on a typical microscope slide. Each slide is hybridized with a fluorescently tagged allele-specific oligonucleotide under TMAC conditions analogous to those described above. The genotype of each individual is determined by the presence or absence of a signal for a selected set of allele-specific oligonucleotides. A schematic of the method is shown in Figure 4.

PCR products are attached to the slide using any methods for attaching DNA to a surface that are known in the art. For instance, PCR products may be spotted onto poly-L-lysine-coated glass slides, and crosslinked by UV irradiation prior to hybridization. A second, more preferred method, which has been developed according to the invention, involves use of oligonucleotides having a 5′ amino group for each of the PCR reactions described above. The PCR products are spotted onto silane-coated slides in the presence of NaOH to covalently attach the products to the slide. This method is advantageous because a covalent bond is formed, which produces a stable attachment to the surface.

SNP-ASO are hybridized under TMAC hybridization conditions with the RCGs covalently conjugated to the surface. The allele-specific oligonucleotides are labeled at their 5'-ends with a fluorescent dye, (e.g., Cy3). After washing, detection of the fluorescent oligonucleotides is performed in one of two ways. Fluorescent images can be captured using a fluorescence microscope equipped with a CCD camera and automated stage capabilities. Alternatively, the data can be obtained using a microarray scanner (e.g. one made by Genetic Microsystems). A microarray scanner provides image analysis which can be converted to a digital (e.g. +/-) signal for each sample using any of several available software applications (e.g., NIH image, ScanAnalyze, etc.). The high signal/noise ratio for this analysis allows for the determination of data in this mode to be straightforward and automated. These data, once exported, can be manipulated to conform with a format which can be analyzed by any of several human genetics applications such as CRI-MAP and LINKAGE software. Additionally, the methods may involve use of two or more fluorescent dyes or other labels which can be spectrally

differentiated to reduce the number of samples which need to be analyzed. For instance, if four fluorescent spectrally distinct dyes, (e.g., ABI Prism dyes 6-FAM, HEX, NED, ROX) are used, then four hybridization reactions can be performed in a single hybridization mixture..

Example 6: Reduction of genome complexity using IRS-PCR or DOP-PCR.

The initial step of the SNP identification method and the genotyping approach described above is to reduce the complexity of genomic DNA in a reproducible manner. The purpose of this step with respect to genotying is to allow genotyping of multiple SNPs using the products of a single PCR reaction. Using the IRS-PCR approach, a PCR primer was synthesized which bears homology to a repetitive sequence present within the genome of the species to be analyzed (e.g., Alu sequence in humans). When two repeat elements bearing the primer sequence are present in a head-to-head fashion within a limited distance (approximately 2 kilobase pairs), the inter-repeat sequence can be amplified. The method has the advantage that the complexity of the resultant PCR can be controlled by how closely the nucleotide sequence primer chosen is to the consensus nucleotide sequence of the repeat element (that is, the closer to the repeat consensus, the more complex the PCR product).

In detail, a 50 microliter reaction for each sample was set up as follows:

distilled, deionized H_20 (ddH_20)	30.75
10X PCR Buffer	5 μl
(500mM KCl, 100mM Tris-HCl pH 8.3, 15mM $MgCl_2$ μM, 0.1% gelatin)	
1.25 mM dNTPs	7.5 μl
20 μm Primer 8C	1.5 μl
Taq polymerase (1.25 units)	0.25 μl
Template (50 ng genomic DNA in ddH_20)	5.0 μl
	50 ul total

The PCR reaction was performed, for example, in a Perkin Elmer 9600 thermal cycler under the following conditions:

1 min.	94°C
30 sec.	94°C |
45 sec.	58°C |32 cycles
90 sec.	72°C |

10 min.	72°C
Hold	4°C

An aliquot of the reaction mixture was separated on an agarose gel to confirm successful amplification.

RCGs were also performed using DOP-PCR with the following primer (CTC GAG NNN NNN AAG CGA TG) (SEQ ID NO: 4) (wherein N is any nucleotide). DOP-PCR uses a single primer which is typically composed of 3 parts, herein designated tag-$(N)_x$-TARGET. The TARGET portion is a polynucleotide which comprises at least 7, and preferably at least 8, arbitrarily-selected nucleotide residues, x is an integer from 0 to 9, and N is any nucleotide residue. Tag is a polynucleotide as described above.

The initial rounds of DOP-PCR were performed at a low temperature, because the specificity of the reaction is determined primarily by the nucleotide sequence of the TARGET portion and the N_x residues. A slow ramp time during these cycles insures that the primers do not detach from the template prior to chain extension. Subsequent amplification rounds were carried out at a higher annealing temperature because of the fact that the 5′ end of the DOP-PCR primer can also contribute to primer annealing.

The DOP-PCR method was performed using a reaction mixture comprising the following ingredients:

distilled deionized H_20	24 μl
10X PCR Buffer	5 μl
1.25 mM dNTPs	8 μl
20 μM Primer DOP-BJ1 (SEQ ID No. 4)	7.5 μl
Taq polymerase	0.5 μl
(1.25 units)	
Template	5 μl
(50 ng genomic DNA in distilled deionized H_20)	50 μl

The PCR reaction was performed, for example, in a Perkin Elmer 9600 thermal cycler using the following reaction conditions:

1 min.	94°C	
1 min.	94°C	|
1.5 min.	45°C	|5 cycles

	2 min. ramp to	72°C |
	3 min.	72 °C|
	1 min.	94°C |
	1.5 min.	58°C |35 cycles
	3 min.	72°C |
	10 min.	72°C
	Hold	4°C

Example 7: Attachment of PCR products to a solid support.

Once the complexity of the genomic DNA from an individual has been reduced, it can be attached to a solid support in order to facilitate hybridization analysis. One method of attaching DNA to a solid support involves spotting PCR products onto a nylon membrane. This protocol was performed as follows:

Upon completion of the PCR reaction (typically in a 50 μl reaction mixture), a 10-fold amount of denaturing solution (500 mM NaOH, 2.0M NaCl, 25 mM EDTA) and a small amount (5 ul) of India Ink were added. Sixty microliters of product was applied to a pre-wetted Hybond ™ N+ membrane (Amersham) using a Schleicher and Schull 96-well dot blot apparatus. The membrane was immediately removed and placed DNA side up on top of Whatmann 3MM paper saturated with 2X SSC for 2 minutes. The filters were air-dried and the DNA was fixed to the membrane by baking in an 80°C oven for 2 hours. The membranes were then used for hybridization.

Another method for attaching nucleic acids to a support involves the use of microarrays. This method attaches minute quantities of PCR products samples onto a glass slide. The number of samples that can be spotted is greater than 1000/cm^2, and therefore over 10,000 samples can be analyzed simultaneously on a glass slide. To accomplish this, pre-cleaned glass slides were placed in a mixture of 80 ml dry xylene, 32 ml 96% 3-glycidoxy-propyltrimethoxy silane, and 160 µl 99% N-ethyldiisopropylamin at 80°C overnight. The slides were rinsed for 5 minutes in ethylacetate and dried at 80°C for 30 minutes. An equal volume of 0.8 M NaOH (0.6M NaOH and 0.6-0.8M KOH also works) was added directly to the PCR product (which contained a 5′ amino group incorporated into the PCR primer) and the components were mixed. The resulting

solution was spotted onto a glass slide under humid conditions. At the earliest opportunity, the slide was placed in a humid chamber overnight at 37°C. The next day, the slide was removed from the humid chamber and kept at 37°C for an additional 1 hour. The slide was incubated in an 80°C oven for 2.5 hours, and then washed for 5 minutes in 0.1% SDS. The slide was washed for an additional 5 minutes in ddH20 and air dried. Attachment to the slide was monitored by OilGreen staining (obtained from Molecular Probes), which specifically binds single-stranded DNA.

Example 8: Hybridization using allele specific oligonucleotides for each SNP.

In order to determine the genotype of an individual at a selected SNP locus, we employed allele-specific oligo hybridizations. Using this method, 2 hybridization reactions were performed at each locus. The first hybridization reaction involved a labeled (radioactive or fluorescent) SNP-ASO (typically 17 nucleotides residues) centered around and complementary to one allele of the SNP. To increase specificity, a 20 to 50-fold excess of non-labeled SNP-ASO complementary to the opposite allele of the SNP was included in the hybridization mixture. For the second hybridization, the allele specificity of the previously labeled and non-labeled SNP-ASOs was reversed. Hybridization occurred in the presence of TMAC buffer, which has the property that oligonucleotides of the same length have the same annealing temperature.

Specifically, for analysis of each SNP, a pair of SNP allele-specific oligos (SNP-ASOs) consisting of two 17mers centered around the polymorphic nucleotide were synthesized. Each filter was hybridized with 20 pmol ^{33}P-labeled kinase labeled SNP-ASO (0.66 pmol/ml) and a 50-fold excess of non-labeled competitor oligonucleotide complementary to the other allele of the SNP. Hybridizations was performed overnight at 52°C in 10 ml TMAC buffer (3.0M TMAC, 0.6% SDS, 1 mM EDTA, 10 mM $NaPO_4$ 6.8, 5X Denhardt's solution, 40 μg/ml yeast RNA). Blots were washed for 20 minutes at room temperature in TMAC Wash Buffer (3M TMAC, 0.6% SDS, 1 mM EDTA, 10 mM Na_3PO_4 pH 6.8) followed by 20 minutes washing at 52°C. The blots were exposed to Kodak X-OMATAR X-ray film for 8-24 hours, and genotypes were determined by analyzing the hybridization pattern.

Example 9: Scoring the hybridization pattern for each sample to determine genotype

Hybridization of SNP-ASOs (2 for each locus) to with IRS-PCR or DOP-PCR products of several individuals has been performed. The final step in this process is to determine if a positive or negative signal exists for each hybridization for an individual and then, based on this information, determine the genotype for that particular locus. Essentially, all of the detection methods described herein can be reduced to a digital image file, for example using a microarray reader or using a phosphoimager. Presently, there are several software products which will overlay a grid onto the image and determine the signal strength value at each element of the grid. These values are imported into a spreadsheet program, like Microsoft Excel™, and simple analysis is performed to assign each signal a + or - value. Once this is accomplished, an individual's genotype can be determined by its pattern of hybridization to the SNP alleles present at a given loci.

Example 10: Genomic Analysis Using DOP-PCR

Genomic DNA isolated from approximately 40 individuals was subjected to DOP-PCR using primer BJ1 (CTC GAG NNN NNN AAG CGA TG) (SEQ ID NO: 4). 100 microliter of the DOP-PCR mixture was precipitated by addition of 10 microliters 3M sodium acetate (pH 5.2) and 110 microliters of isopropanol and were stored at -20°C for at least 1 hour. The samples were spun down in a microcentrifuge for 30 minutes and the supernatant was removed. The pellets were rinsed with 70% ethanol and spun again for 30 minutes. The supernatant was removed and the pellets were air-dried overnight at room temperature.

The pellets were then resuspended in 12 microliters of distilled water and stored at -20°C until denatured by the addition of 3 microliter of 2N NaOH/50 mM EDTA and maintained at 37°C for 20 minutes and then at room temperature for 15 minutes. The samples were then spotted onto nylon coated-glass slides using a Genetic Microsystems GMS417 microarrayer. Upon completion of the spotting, the slides were placed in an 80 °C vacuum oven for 2 hours, and then stored at room temperature. A set of 2 allele specific SNP-ASOs consisting of two 17mers centered around a polymorphic nucleotide residue were synthesized. Each slide was prehybridized for 1 hour in Hyb Buffer (3M TMAC/0.5% SDS/1mM EDTA/10 mM $NaPO_4$/5X

Denhardt's solution/40 μg/ml yeast RNA) followed by hybridization with .66 picomoles per milliliter ^{33}P-labeled kinase labeled SNP-ASO and a 50- fold excess of cold-competitor SNP-ASO of the opposite allele in Hyb Buffer. Hybridizations were carried out overnight at 52°C. The slides were washed twice for 30 minutes at room temperature in TMAC Wash Buffer (3M TMAC, 0.6% SDS, 1 mM EDTA, 10 mM $NaPO_4$ pH 6.8) followed by 20 minutes at 54°C. The slides were exposed to Kodak BioMax MR X-ray film. The results are shown in Figure 8. The genotypes were determined by the hybridization patterns shown in Figure 8 wherein loci are indicated.

The foregoing written specification is considered to be sufficient to enable one skilled in the art to practice the invention. The present invention is not limited in scope by the examples provided, since the examples are intended as illustrations of various aspect of the invention and other functionally equivalent embodiments are within the scope of the invention. Various modifications of the invention in addition to those shown and described herein will become apparent to those skilled in the art from the foregoing description and fall within the scope of the appended claims. The advantages and objects of the invention are not necessarily encompassed by each embodiment of the invention.

All references, patents and patent publications that are recited in this application are incorporated in their entirety herein by reference.

We claim:

CLAIMS

1. A method for detecting the presence or absence of a single nucleotide polymorphism (SNP) allele in a genomic sample, the method comprising:

preparing a reduced complexity genome (RCG) from the genomic sample, and

analyzing the RCG for the presence or absence of a SNP allele.

2. The method of claim 1, wherein the analysis comprises hybridizing a SNP-ASO and the RCG, wherein the SNP-ASO is complementary to one allele of a SNP, whereby the allele of the SNP is present in the genomic sample if the SNP-ASO hybridizes with the RCG, and wherein the presence or absence of the SNP is used to characterize the genomic sample.

3. The method of claim 2, wherein the RCG is immobilized on a surface.

4. The method of claim 2, wherein the SNP-ASO is immobilized on a surface.

5. The method of claim 2, wherein the SNP-ASO is individually hybridized with a plurality of RCGs.

6. The method of claim 1, wherein the RCG is a PCR-derived RCG.

7. The method of claim 1, wherein the RCG is a native RCG.

8. The method of any one of claims 1-7, wherein the method further comprises identifying a genotype of the genomic sample, whereby the genotype is identified by the presence or absence of the alleles of the SNP in the RCG.

9. The method of any one of claims 1-7, wherein the genomic sample is obtained from a tumor.

10. The method of claim 9, wherein a plurality of RCGs are prepared from genomic samples isolated from a plurality of subjects and the plurality of RCGs are analyzed for the presence of the SNP.

11. The method of claim 8, wherein the presence or absence of the SNP allele is analyzed in a plurality of genomic samples selected randomly from a population, the method further comprising determining the allelic frequency of the SNP allele in the population by comparing the number of genomic samples in which the allele is detected and the number of genomic samples analyzed.

12. The method of claim 1, wherein the RCG is prepared by performing degenerate oligonucleotide priming-polymerase chain reaction (DOP-PCR) using a degenerate oligonucleotide primer having a tag-$(N)_x$-TARGET nucleotide sequence, wherein the TARGET nucleotide sequence includes at least 7 TARGET nucleotide residues, wherein x is an integer from 0-9, and wherein each N is any nucleotide residue, and wherein the tag is a polynucleotide having from about 0 to about 20 nucleotides.

13. The method of claim 12, wherein the TARGET nucleotide sequence includes at least 8 nucleotide residues.

14. The method of claim 6, wherein the RCG is prepared by interspersed repeat sequence-polymerase chain reaction (IRS-PCR).

15. The method of claim 6, wherein the RCG is prepared by arbitrarily primed-polymerase chain reaction (AP-PCR).

16. The method of claim 6, wherein the RCG is prepared by adapter-polymerase chain reaction.

17. The method of claim 2, wherein at least a fraction of the SNP-ASO is labeled.

18. The method of claim 17, wherein an excess of a non-labeled SNP-ASO is added during the hybridization step, wherein the non-labeled oligonucleotide is complementary to a different allele of the same SNP than the labeled SNP-ASO.

19. The method of claim 17, further comprising performing a parallel hybridization reaction wherein the RCG is hybridized with a labeled SNP-ASO, wherein the oligonucleotide is complementary to a different allele of the same SNP than the labeled SNP-ASO.

20. The method of claim 19, wherein the two SNP-AGOs are distinguishably labeled.

21. The method of claim 17, an excess of non-labeled SNP-ASO is present during the hybridization.

22. The method of claim 2, wherein the SNP-ASO is composed of from about 10 to about 50 nucleotides residues.

23. The method of claim 22, wherein the SNP-ASO is composed of from about 10 to about 25 nucleotides residues.

24. The method of claim 17, wherein the label is a radioactive isotope.

25. The method of claim 24, further comprising the step of exposing the RCG to a film to produce a signal on the film which corresponds to the radioactively labeled hybridization products if the SNP is present in the RCG.

26. The method of claim 17, wherein the label is a fluorescent molecule.

27. The method of claim 26, further comprising the step of exposing the RCG to an automated fluorescence reader to generate an output signal which corresponds to the fluorescently labeled hybridization products if the SNP is present in the RCG.

28. The method of claim 17, wherein a plurality of SNP-ASOs are labeled with fluorescent molecules, each SNP-ASO being labeled with a spectrally distinct fluorescent molecule.

29. The method of claim 28, wherein the number of SNP-ASOs having a spectrally distinct fluorescent molecule is at least two.

30. The method of claim 28, wherein the number is selected from the group consisting of three, four and eight.

31. The method of claim 2, wherein a plurality of RCGs are labeled with fluorescent molecules, each RCG being labeled with a spectrally distinct fluorescent molecule, and wherein all of the RCGs having a spectrally distinct fluorescent molecule.

32. The method of claim 1, wherein the RCG is prepared by performing degenerate oligonucleotide priming-polymerase chain reaction using a degenerate oligonucleotide primer having a tag-$(N)_x$-TARGET nucleotide sequence, wherein the TARGET nucleotide sequence includes fewer than 7 TARGET nucleotide residues wherein x is an integer from 0 to 9, wherein each N is any nucleotide residues, and wherein the tag is a polynucleotide having from about 0-20 nucleotides.

33. The method of claim 32 wherein the TARGET nucleotide sequence includes at least 5 nucleotide residues.

34. The method of claim 32 wherein the TARGET nucleotide sequence includes at

least 6 nucleotide residues.

35. The method of claim 2, wherein the RCG is labeled.

36. The method of claim 4, wherein a plurality of different SNP-ASOs are attached to the surface.

37. The method of claim 1, wherein the RCG is prepared by performing multiple primed DOP-PCR.

38. The method of claim 2, wherein the genomic sample is characterized by generating a genomic pattern based on the presence or absence of the allele of the SNP in the genomic sample.

39. The method of claim 38, wherein the genomic pattern is a genomic classification code.

40. A method for characterizing a tumor, the method comprising:

isolating genomic DNA from tumor samples obtained from a plurality of subjects,

preparing a RCGs from each genomic DNA,

performing a hybridization reaction with a SNP-ASO and the plurality of RCGs, wherein the SNP-ASO is complementary to one allele of a SNP, and

characterizing the tumor based on whether the SNP-ASO hybridizes with at least some of the RCGs, whereby if the SNP oligonucleotide hybridizes with at least some of the RCGs, then the allele of the SNP is present in the genomic DNA of the tumor.

41. A method for generating a genomic pattern for an individual genome, the method comprising:

preparing a RCG from the individual genome,

analyzing the RCG for the presence or absence of at least one SNP allele, and

generating a genomic pattern for the individual genome based on the presence or absence of SNP alleles.

42. The method of claim 41, wherein analyzing the RCG involves a hybridizing the RCG with a panel of SNP-ASOs, each of which is complementary to one allele of a SNP, and identifying the genomic pattern by determining the ability of the RCG to hybridize with each SNP-ASO.

43. The method of claim 41, wherein the genomic pattern is a genomic identification code which is generated from the pattern of SNP alleles for each RCG.

44. The method of claim 43, wherein the genomic classification code is also generated using the allelic frequency of the SNPs.

45. The method of claim 41, wherein the genomic pattern is a visual pattern.

46. The method of claim 41, wherein the genomic pattern is a digital pattern.

47. A method for generating a genomic classification code for a genome, the method comprising:

preparing a RCG from the genome,

analyzing the RCG for the presence or absence of SNP alleles of known allelic frequency, and

identifying a genomic pattern of SNP alleles for the RCG by determining the presence or absence therein of SNP alleles, and

generating a genomic classification code for the RCG based on the presence or absence and the allelic frequency of the SNP alleles.

48. A composition, comprising:

a plurality of RCGs immobilized in an ordered array on a surface.

49. The composition of claim 48, wherein the RCGs prepared by the method of claim 125.

50. The composition of claim 48, wherein the RCGs are PCR-generated RCGs.

51. The composition of claim 48, wherein the RCGs are native RCGs.

52. A kit, comprising:

a container housing a set of polymerase chain reaction primers for reducing the complexity of a genome, and

a container housing a set of SNP-ASOs, wherein the SNPs are present with a frequency of at least 50% in a RCG made using the set of primers.

53. The kit of claim 52, wherein the SNP-ASOs are attached to a surface.

54. The kit of any one of claims 52 or 53, wherein the set of polymerase chain reaction primers are primers for DOP-PCR.

55. The kit of claim 54, wherein the degenerate oligonucleotide primer has a tag-$(N)_x$-TARGET nucleotide sequence, wherein the TARGET nucleotide sequence includes at least 7 TARGET nucleotide residues and wherein x is an integer from 0 to 9, wherein each N is any nucleotide residue, and wherein each tag is a polynucleotide having from 0 to about 20 nucleotide residues.

56. The kit of claim 55, wherein the TARGET nucleotide sequence includes at least 8 nucleotide residues.

57. The kit of any one of claims 52 or 53, wherein the SNP-ASOs are composed from 10 and 50 nucleotide residues.

58. The kit of any one of claims 52 or 53, wherein the SNP-ASOs are composed of from 10 and 25 nucleotide residues.

59. The kit of claim 54, wherein the degenerate oligonucleotide primer has a tag-$(N)_x$-TARGET nucleotide sequence, wherein the TARGET nucleotide sequence includes fewer than 7 TARGET nucleotide residues and wherein x is an integer from 0 to 9, wherein each N is any nucleotide residue, and wherein each tag is a polynucleotide having from 0 to about 20 nucleotide residues.

60. The kit of claim 52, wherein the set of polymerase chain reaction primers are primers for multiple-primed DOP-PCR.

61. A composition comprising:

a plurality of RCGs immobilized on a surface, wherein the RCGs are composed of a plurality of DNA fragments, each DNA fragment comprising a $(N)_x$-TARGET nucleotide portion, wherein the nucleotide sequence of TARGET is identical in each of the DNA fragments, wherein TARGET is a polynucleotide consisting of at least 7 nucleotide residues, wherein x is an integer from 0 to 9, and wherein N is any nucleotide residue.

62. The composition of claim 61, wherein the TARGET nucleotide sequence includes 8 nucleotide residues.

63. A method for identifying a SNP, the method comprising:

preparing a set of primers from a RCG, wherein the RCG comprises a set of polymerase chain reaction (PCR) products,

performing PCR using the set of primers on at least one of isolated genome to produce a set of DNA products, and

identifying a SNP on the set of DNA products.

64 The method of claim 63, wherein the plurality of isolated genomes is a pool of genomes.

65. The method of claim 63,, wherein the isolated genomes are RCGs.

66. The method of claim 65, wherein the RCG is prepared by DOP-PCR.

67. The method of claim 63, wherein the step of preparing the set of primers is performed by at least the following steps:

preparing a RCG and separating the set of PCR products in the RCG into individual PCR products,

determining the sequence of each end of at least one of the PCR products, and

generating primers for use in the subsequent PCR step based on the sequence of the ends of the inserts.

68. The method of claim 63, wherein the RCG is prepared by performing DOP-PCR using a degenerate oligonucleotide primer having a tag-$(N)_x$-TARGET nucleotide sequence, wherein the TARGET nucleotide sequence includes at least 7 TARGET nucleotide residues and wherein x is an integer from 0 to 9, wherein each N is any nucleotide residue, and wherein each tag is a polynucleotide having from 0 to about 20 nucleotide residues.

69. The method of claim 68, wherein the TARGET nucleotide sequence includes 8 nucleotide residues.

70. The method of claim 63, wherein the RCG is prepared by performing DOP-PCR using a degenerate oligonucleotide primer having a tag-$(N)_x$-TARGET nucleotide sequence, wherein the TARGET nucleotide sequence includes less than 7 TARGET nucleotide residues and wherein x^1 is an integer from 0 to 9, wherein each N is any nucleotide residue, and wherein each tag is a polynucleotide having from 0 to about 20 nucleotide residues.

71. A composition comprising:

a panel of SNP-ASOs immobilized on a surface, wherein the SNP-ASOs are prepared by the method of claim 63.

72. The composition of claim 71, wherein each SNP-ASO is immobilized in a discrete area of the surface.

73. The composition of claim 71, further comprising a panel of complementary SNP-ASOs immobilized on discrete areas of the surface.

74. A method for obtaining a RCG using DOP-PCR, the method comprising:

performing DOP-PCR using a degenerate oligonucleotide primer having a tag-$(N)_x$-TARGET nucleotide sequence, wherein the TARGET nucleotide sequence includes at least 7 TARGET nucleotide residues and wherein x is an integer from 0 to 9, wherein each N is any nucleotide residue, and wherein each tag is a polynucleotide having from 0 to about 20 nucleotide residues.

75 The method of claim 74, wherein the TARGET nucleotide sequence includes 8 nucleotide residues.

76. The method of any one of 74-75, further comprising using the RCG in a genotyping procedure.

77. The method of any one of 74-75, further comprising analyzing the RCG to detect a polymorphism.

78. The method of claim 77 wherein the RCG is analyzed using mass spectroscopy.

79. A method for assessing whether a subject is at risk for developing a disease, the method comprising:

preparing a RCG from a genomic sample obtained from the subject and characterizing the sample by the method of claim 1, whether one sample based on the presence or absence in the sample of a plurality of SNP alleles that occur in at least 10% of genomes obtained from individuals afflicted with the disease occur in the reduced subject complexity genome.

80. A method for identifying a set of SNP alleles associated with a disease, the method comprising:

preparing individual RCGs obtained from subjects afflicted with a disease using the same set of primers to prepare each RCG, and

comparing individual genetic loci in the RCGs with the same individual genetic loci in normal subjects to identify SNP associated with the disease.

81. A digital information product for representing genomic information, the product comprising:

a computer-readable medium having computer-readable signals stored thereon, wherein the signals define a data structure, the data structure including one or more data components,

wherein each data component includes:

a first data element defining a genomic classification code that identifies a corresponding genome, and wherein each genomic classification code classifies the corresponding genome based one or more single nucleotide polymorphisms of the corresponding genome.

82. The digital information produce of claim 81, wherein the genomic classification code is a unique identifier of the corresponding genome.

83. The digital information product of claim 81, wherein the genomic classification code is based on a pattern of the single nucleotide polymorphisms of the corresponding genome, the pattern indicating the presence or absence of each single nucleotide polymorphism.

84. The digital information product of claim 81, wherein each data component also includes:

one or more data elements, each data element defining an attribute of the corresponding genome.

85. A process for making a digital information product comprising computer data signals defining a genomic classification code for a genome, the process comprising:

preparing a reduced complexity genome,

performing a hybridization reaction with the reduced complexity genome and at least one surface having a panel of single nucleotide polymorphism oligonucleotides immobilized thereon,

identifying a genomic pattern of single nucleotide polymorphisms for the reduced complexity genome by determining the presence therein of single nucleotide polymorphisms based on whether each single nucleotide polymorphism oligonucleotide hybridizes to the reduced complexity genome,

generating a genomic classification code for the reduced complexity genome based on the genomic pattern of the single nucleotide polymorphisms, and

encoding the genomic classification code as one or more computer data signals on a computer-readable medium.

86. A process for making a digital information product comprising computer data signals defining a genomic classification code for a genome, the process comprising:

preparing a reduced complexity genome,

performing a hybridization reaction with a panel of single nucleotide polymorphism oligonucleotides of known allelic frequency and a surface having the reduced complexity genome immobilized thereon,

identifying a genomic pattern of single nucleotide polymorphisms for the reduced complexity genome by determining the presence therein of single nucleotide polymorphisms based on whether each single nucleotide polymorphism oligonucleotide hybridizes to the reduced complexity genome,

generating a genomic classification code for the reduced complexity genome based on the pattern and the allelic frequency of the single nucleotide polymorphisms, and

encoding the genomic classification code as one or more computer data signals on a computer-readable medium.

87 A method for performing linkage analysis, comprising:

preparing individual RCGs obtained from members of one or more families ,

determining the presence or absence of SNP alleles in the RCGs, and

comparing the RCGs of the family members by comparing the presence or absence of the SNP alleles in the RCGs of the family members.

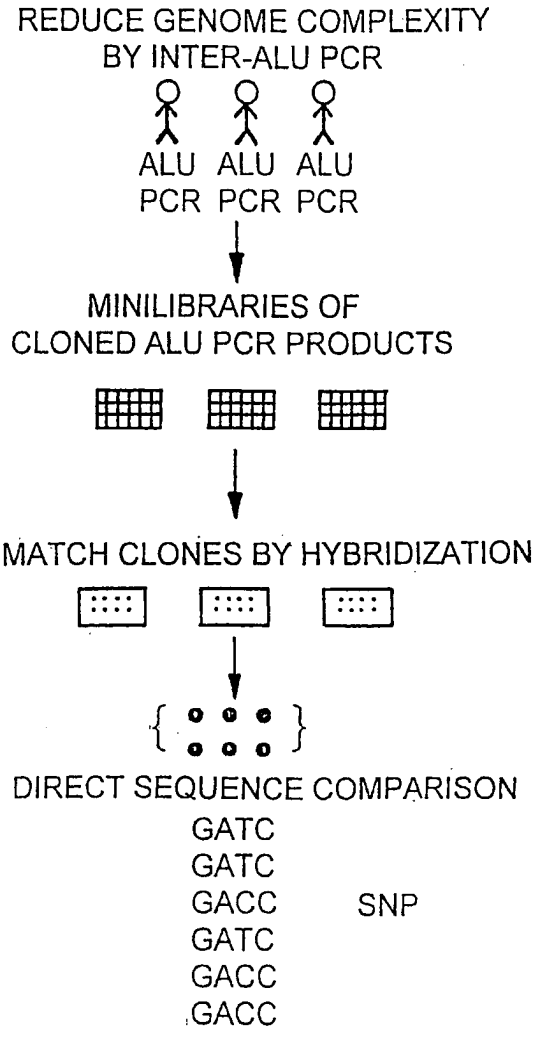
REDUCE GENOME COMPLEXITY
BY INTER-ALU PCR
ALU ALU ALU
PCR PCR PCR
MINILIBRARIES OF
CLONED ALU PCR PRODUCTS
MATCH CLONES BY HYBRIDIZATION
DIRECT SEQUENCE COMPARISON
GATC
GATC
GACC
GATC
GACC
GACC
SNP

Fig. 1

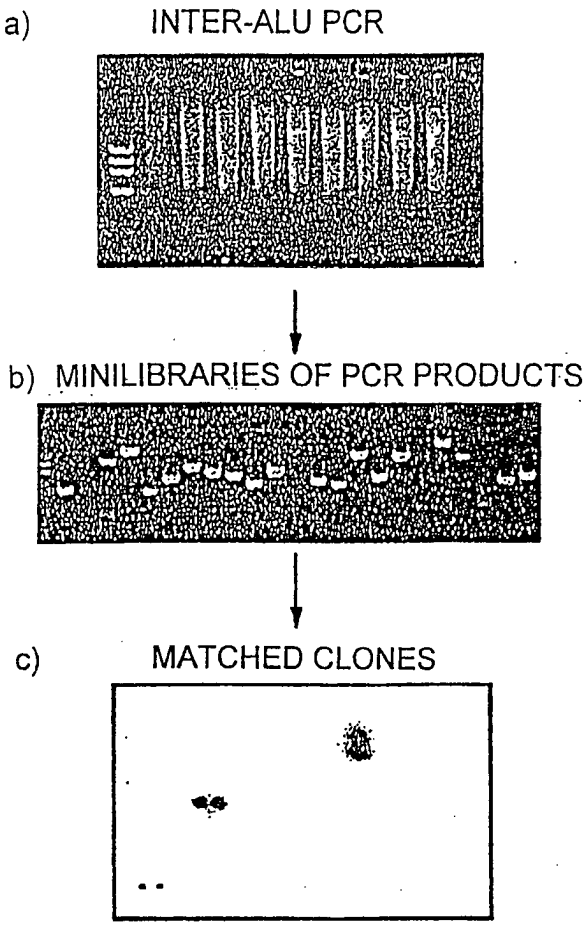
a)
INTER-ALU PCR
b) MINILIBRARIES OF PCR PRODUCTS
c)
MATCHED CLONES

Fig. 2

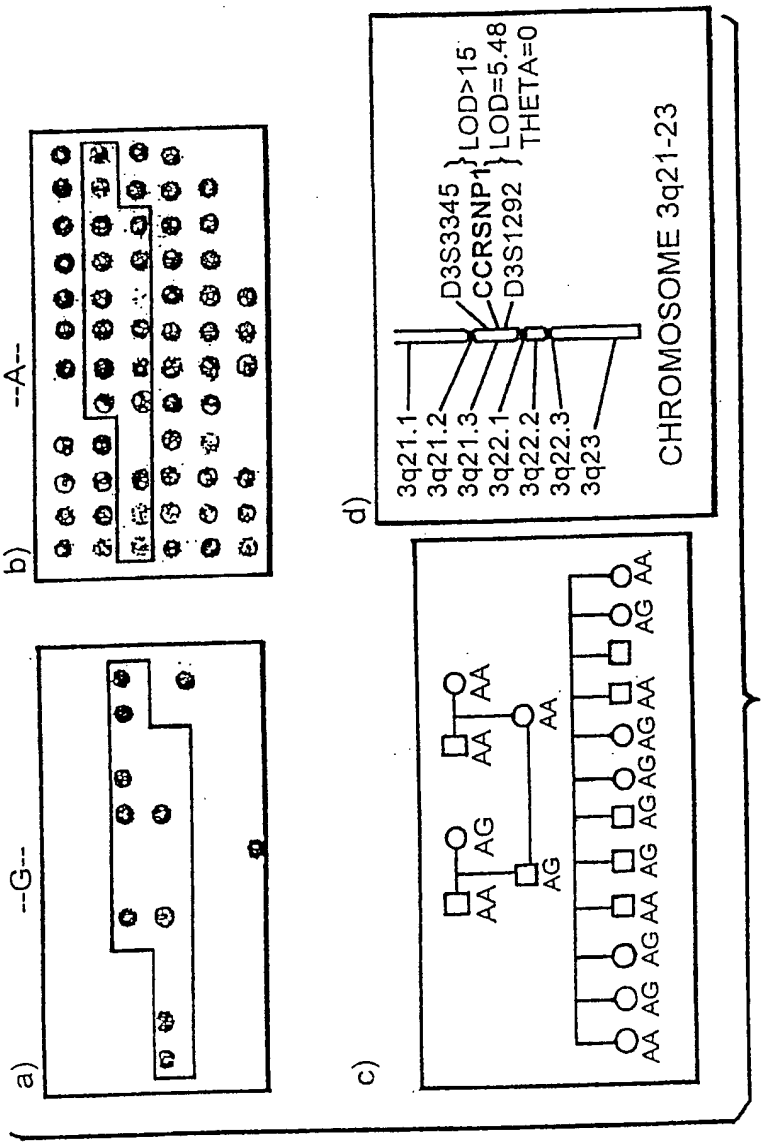
a)
--G--
b)
--A--
c)
AA
AG
AA
AA
AG
AA
AG
AG
AG
AA
AG
AA
AG
AA
AA
d)
3q21.1
3q21.2
3q21.3
3q22.1
3q22.2
3q22.3
3q23
D3S3345
CCRSNP1
D3S1292
LOD>15
LOD=5.48
THETA=0
CHROMOSOME 3q21-23

Fig. 3

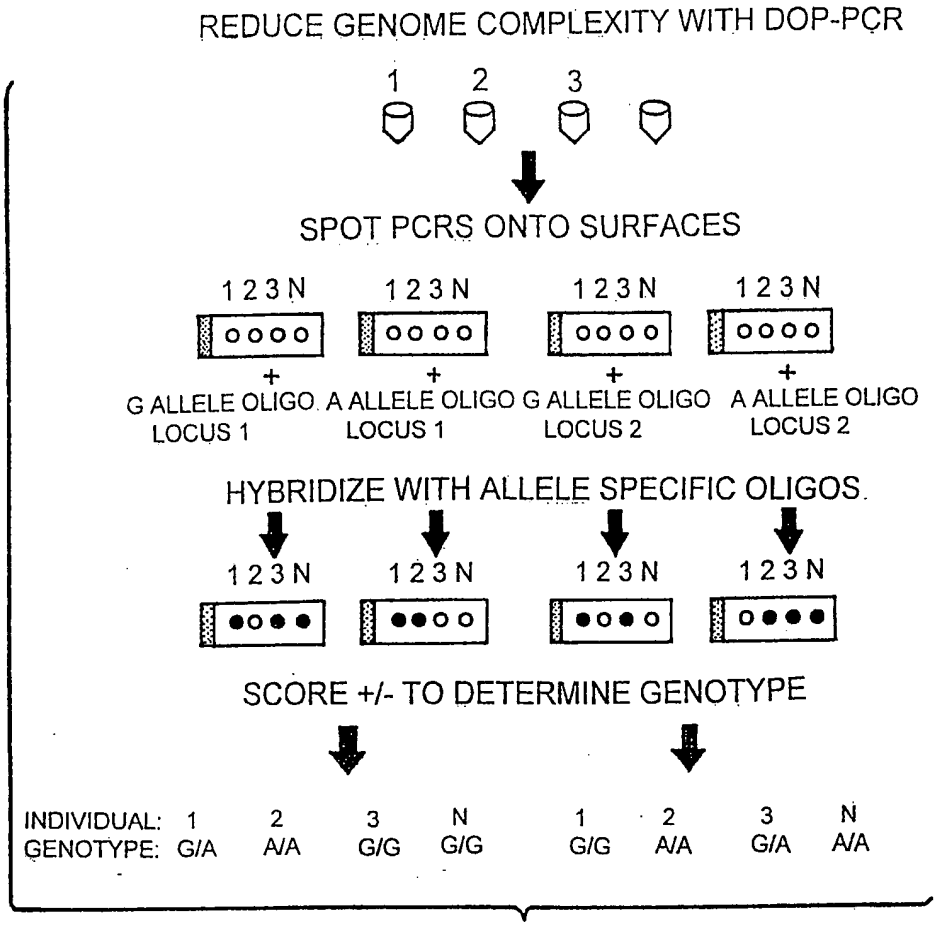
REDUCE GENOME COMPLEXITY WITH DOP-PCR
1 2 3
SPOT PCRS ONTO SURFACES
1 2 3 N
1 2 3 N
1 2 3 N
1 2 3 N
+
+
+
+
G ALLELE OLIGO LOCUS 1
A ALLELE OLIGO LOCUS 1
G ALLELE OLIGO LOCUS 2
A ALLELE OLIGO LOCUS 2
HYBRIDIZE WITH ALLELE SPECIFIC OLIGOS
1 2 3 N
1 2 3 N
1 2 3 N
1 2 3 N
SCORE +/- TO DETERMINE GENOTYPE
INDIVIDUAL: 1 2 3 N 1 2 3 N
GENOTYPE: G/A A/A G/G G/G G/G A/A G/A A/A

Fig. 4

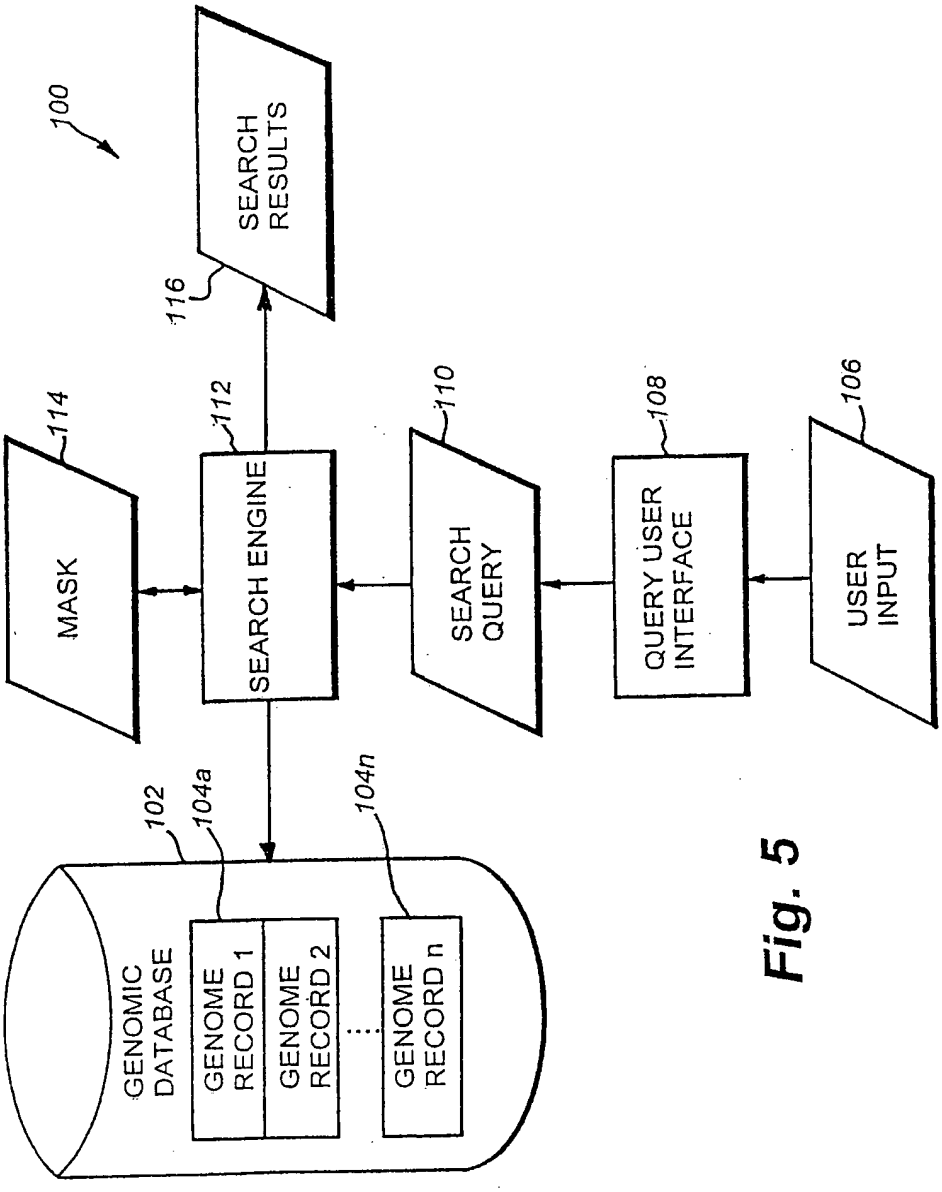
100
SEARCH RESULTS
116
SEARCH ENGINE
112
MASK
114
SEARCH QUERY
110
QUERY USER INTERFACE
108
USER INPUT
106
GENOMIC DATABASE
102
GENOME RECORD 1
104a
GENOME RECORD 2
GENOME RECORD n
104n

Fig. 5

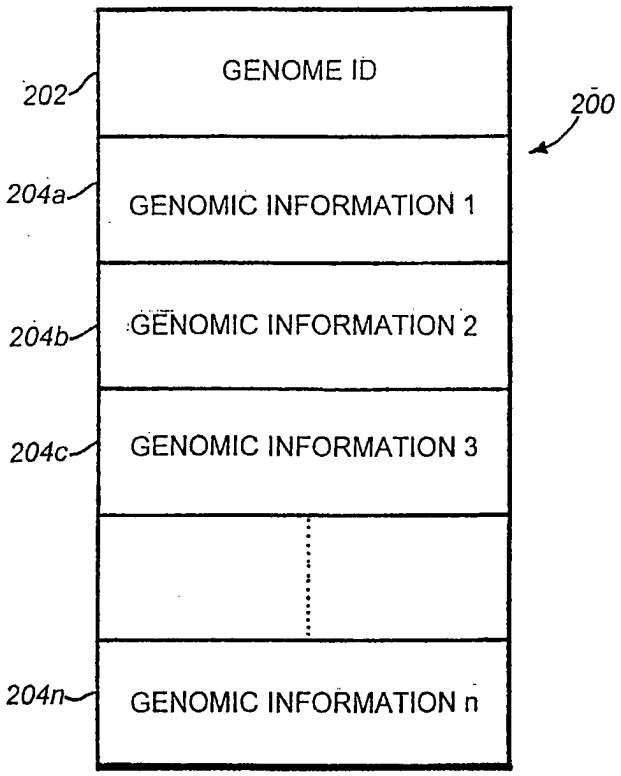
GENOME ID
202
200
204a
GENOMIC INFORMATION 1
204b
GENOMIC INFORMATION 2
204c
GENOMIC INFORMATION 3
204n
GENOMIC INFORMATION n

Fig. 6A

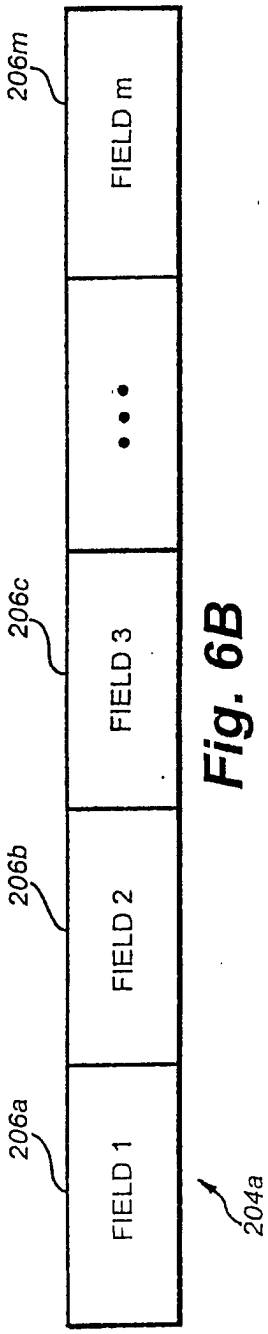

Fig. 6B

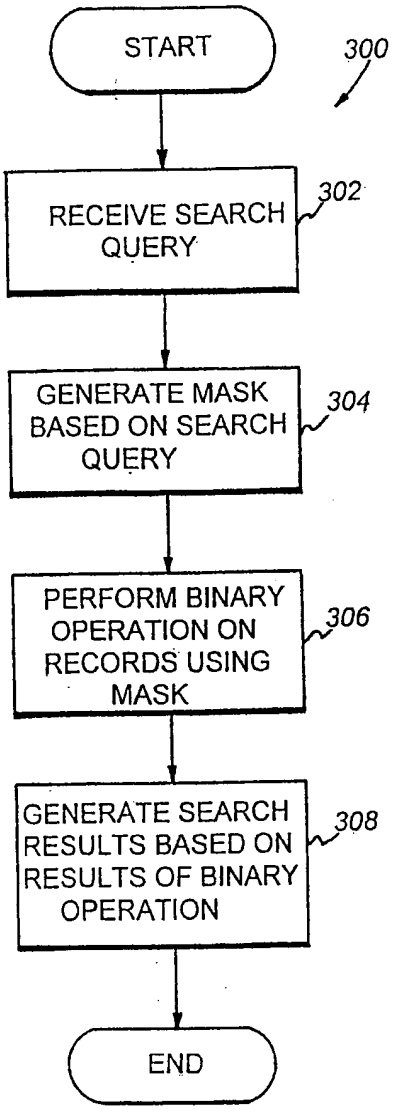
START
300
RECEIVE SEARCH QUERY
302
GENERATE MASK BASED ON SEARCH QUERY
304
PERFORM BINARY OPERATION ON RECORDS USING MASK
306
GENERATE SEARCH RESULTS BASED ON RESULTS OF BINARY OPERATION
308
END

Fig. 7

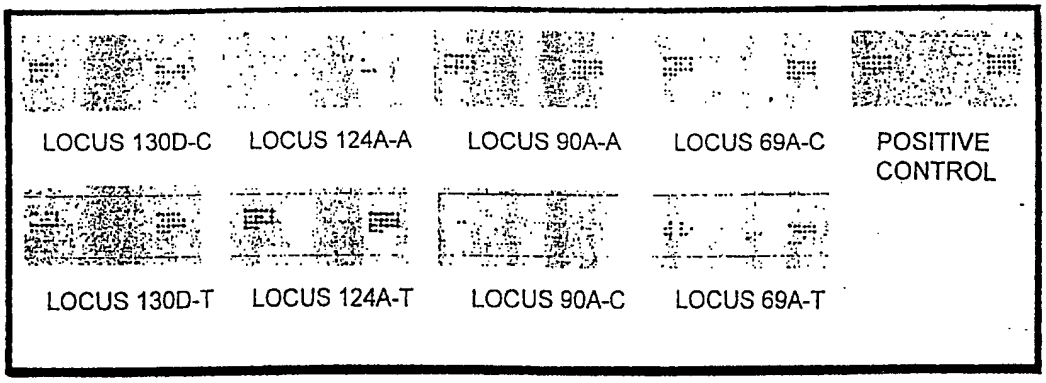
LOCUS 130D-C
LOCUS 124A-A
LOCUS 90A-A
LOCUS 69A-C
POSITIVE CONTROL
LOCUS 130D-T
LOCUS 124A-T
LOCUS 90A-C
LOCUS 69A-T

Fig. 8

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : C12Q 1/68, G06F 17/30	A3	(11) International Publication Number: WO 00/18960
		(43) International Publication Date: 6 April 2000 (06.04.00)

(21) International Application Number: PCT/US99/22283

(22) International Filing Date: 24 September 1999 (24.09.99)

(30) Priority Data:
60/101,757 25 September 1998 (25.09.98) US

(71) Applicant: MASSACHUSETTS INSTITUTE OF TECHNOLOGY [US/US]; 77 Massachusetts Avenue, Cambridge, MA 02139 (US).

(72) Inventors: LANDERS, John, E.; 77 Massachusettts Avenue, Cambridge, MA 02139 (US). JORDAN, Barbara; 77 Massachusetts Avenue, Cambridge, MA 02139 (US). HOUSMAN, David, E.; 77 Massachusetts Avenue, Cambridge, MA 02139 (US). CHAREST, Alain; 77 Massachusetts Avenue, Cambridge, MA 02139 (US).

(74) Agent: LOCKHART, Helen, C.; Wolf, Greenfield & Sacks, P.C., 600 Atlantic Avenue, Boston, MA 02210 (US).

(81) Designated States: AU, CA, IL, IS, JP, NO, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(88) Date of publication of the international search report:
8 September 2000 (08.09.00)

(54) Title: METHODS AND PRODUCTS RELATED TO GENOTYPING AND DNA ANALYSIS

(57) Abstract

The invention encompasses methods and products related to genotyping. The method of genotyping of the invention is based on the use of single nucleotide polymorphisms (SNPs) to perform high throughput genome scans. The high throughput method can be performed by hybridizing SNP allele-specific oligonucleotides and a reduced complexity genome (RCG). The invention also relates to methods of preparing the SNP specific oligonucleotides and RCGs, methods of fingerprinting, determining allele frequency for an SNP, characterizing tumors, generating a genomic classification code for a genome, identifying previously unknown SNPs, and related compositions and kits.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece			TR	Turkey
BG	Bulgaria	HU	Hungary	ML	Mali	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MN	Mongolia	UA	Ukraine
BR	Brazil	IL	Israel	MR	Mauritania	UG	Uganda
BY	Belarus	IS	Iceland	MW	Malawi	US	United States of America
CA	Canada	IT	Italy	MX	Mexico	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NE	Niger	VN	Viet Nam
CG	Congo	KE	Kenya	NL	Netherlands	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NO	Norway	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	NZ	New Zealand		
CM	Cameroon			PL	Poland		
CN	China	KR	Republic of Korea	PT	Portugal		
CU	Cuba	KZ	Kazakstan	RO	Romania		
CZ	Czech Republic	LC	Saint Lucia	RU	Russian Federation		
DE	Germany	LI	Liechtenstein	SD	Sudan		
DK	Denmark	LK	Sri Lanka	SE	Sweden		
EE	Estonia	LR	Liberia	SG	Singapore		

INTERNATIONAL SEARCH REPORT

Inter onal Application No
PCT/US 99/22283

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 C12Q1/68 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	BROUDE ET AL.: "DIFFERENTIAL DISPLAY OF GENOME SUBSETS CONTAINING SPECIFIC INTERSPERSED REPEATS" PNAS, vol. 94, April 1997 (1997-04), pages 4548-4553, XP002138635 the whole document	1-87
Y	CHENG ET AL.: "DEGENERATE OLIGONUCLEOTIDE PRIMER-POLYMERASE CHAIN REACTION AND CAPILLARY ELECTROPHORETIC ANALYSIS OF HUMAN DNA ON MICROCHIP-BASED DEVICES" ANAL.BIOCHEM., vol. 257, March 1998 (1998-03), pages 101-106, XP002138636 the whole document	1-87

-/--

[X] Further documents are listed in the continuation of box C.

[X] Patent family members are listed in annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
30 May 2000	13/06/2000

Name and mailing address of the ISA	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	Hagenmaier, S

1

Form PCT/ISA/210 (second sheet) (July 1992)

INTERNATIONAL SEARCH REPORT

Inte. onal Application No
PCT/US 99/22283

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication,where appropriate, of the relevant passages	Relevant to claim No.
Y	WO 98 20165 A (WHITEHEAD BIOMEDICAL INST ;HUDSON THOMAS (US); LANDER ERIC S (US);) 14 May 1998 (1998-05-14) the whole document	1-87
Y	WANG D ET AL: "Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome" SCIENCE,US,AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE,, vol. 280, no. 280, 15 May 1998 (1998-05-15), pages 1077-1082-82, XP002116081 ISSN: 0036-8075 the whole document	1-87
Y	WO 97 31327 A (MOTOROLA INC ;REBER WILLIAM L (US); PERTTUNEN CARY D (US)) 28 August 1997 (1997-08-28) the whole document	81-86
A	HIMMELBAUER ET AL.: "COMPLEX PROBES FOR HIGH-THROUGHPUT PARALLEL GENETIC MAPPING OF GENOMIC MOUSE BAC CLONES" MAMMALIAN GENOME, vol. 9, August 1998 (1998-08), pages 611-616, XP000907317 the whole document	
A	XIONG M AND JIN L: "Biallelic markers in genetics studies of human diseases: Their power, accuracy, and density in population-based linkage analysis" AMERICAN JOURNAL OF HUMAN GENETICS,US,NEW YORK, NY, vol. 61, no. 4, SUPPL, 1997, page 1759 XP002119235 ISSN: 0002-9297 the whole document	
A	KRUGLYAK L: "THE USE OF A GENETIC MAP OF BIALLELIC MARKERS IN LINKAGE STUDIES" NATURE GENETICS,US,NEW YORK, NY, vol. 17, no. 1, 1 September 1997 (1997-09-01), pages 22-24, XP002050647 ISSN: 1061-4036 the whole document	

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

Inte ional Application No
PCT/US 99/22283

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 9820165 A	14-05-1998	EP 0941366 A	15-09-1999
WO 9731327 A	28-08-1997	AU 1414197 A	10-09-1997